Addressing Time-series Signal Quality in Healthcare Data

Chufan Gao

CMU-RI-TR-22-47

August 18, 2022

The Robotics Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Artur Dubrawski, *chair*
Jeff Schneider
Gilles Clermont (University of Pittsburgh)
Benedikt Boecking

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Robotics.

Abstract

Healthcare data time-series signal quality assessment (SQA) plays a vital role in the accuracy and reliability of machine learning algorithms to analyze health metrics. However, these signals are often corrupted with different kinds of noises and artifacts, including Baseline Wander, Muscle Artifacts, Powerline Interference, and Equipment Failure. This can lead to vital, potentially deadly, errors in the medical domain. This can include inaccurate calculation of basic health features like Heart Rate, clinical alarm burnout from bedside monitors, as well as disrupting general downstream machine learning tasks. While some work has been done in the area of open-source signal quality analysis in general, there are very few open source implementations of signal quality analysis frameworks that attempt to reproduce and expand on existing results on open source datasets.

First, we propose an open-source implementation of signal quality indices (SQIs) for analysis of electrocardiogram (ECG), plethysmography, and more. We aim to codify and reproduce SQIs and results from The Physionet Signal Quality Classification 2011 Challenge. We show that these SQIs may be used for signal quality outlier detection in a real world clinical dataset from University of Pittsburgh Medical Center (UPMC). Secondly, in the case of another common healthcare SQA issue: ECG denoising, we compare Wavelet, EMD, and Convolutional Autoencoder denoising techniques. We show that Convolutional Autoencoder denoising performs the best on the open MIT-BIH Arrhythmia Noise Stress Test dataset, and evaluate it on the UPMC dataset. We also perform a case

study on ECG denoising for real vs artifactual alert classification. To our knowledge, we are the first to provide an open source implementation of these two SQA tasks that is validated on public datasets. Ideally, this work serves as an accessible, open source, toolkit for signal quality analysis and ECG denoising.

Acknowledgments

I would like to acknowledge the wonderful members of the AutonLab that consistently challenges and supports me to become a better researcher, student, and person in general. *Once an Autonian, always an Autonian.* I would like to give extra thanks to Professor Artur Dubrawski, Mononito Goswami, Cecilia Morales, and Xinyu Li for listening to and discussing with me in times of stress and anxiety, which occurred fairly often.

Furthermore, I would like to explicitly thank Rachel Burcin and John Dolan for their gracious and consistent support of the Robotics Institute Summer Scholars (RISS) program, as it changed my life forever.

Finally, to my parents: Thank you.

Contents

When this dissertation is viewed as a PDF, the page header is a link to this Table of Contents.

List of Figures

List of Tables

Chapter 1

Introduction

Healthcare data time-series signal quality assessment (SQA) plays a vital role in the accuracy and reliability of machine learning (ML) algorithms to analyze health metrics. However, these signals are often corrupted with different kinds of noise and artifact such as Baseline Wander, Muscle Artifacts, Powerline Interference, and Equipment Failure. This can lead to vital, potentially deadly, errors in clinical practice. Signal quality is crucial in false alarm reduction [5, 17, 34, 54], physiological pattern discovery [22, 23], modelling electrocardiogram data [27, 46], determining cardiovascular sufficiency [56, 69], and more. In all of these applications, issues with quality can significantly reduce the accuracy and applicability of these ML algorithms.

One of the largest bodies of work in the signal quality space involves Electrocardiogram (ECG) signal quality assessment. Since it includes many approaches that are generalizable to signal quality analysis as a whole, much of this work is dedicated to ECG signal quality. While some work has been done in the area of open-source signal quality analysis, there are very few open source implementations of signal quality analysis frameworks that attempt to reproduce and expand on existing results on open source datasets. For example, Neurokit2 [1][61] - one of the largest ECG analysis libraries - only implements 2 signal quality analysis indices, one of which does not work well in practice. Additionally, there is a lack practical evaluation on open source datasets. Aura Healthcare[2] - another ECG analysis repository - includes methods

[1] github.com/neuropsychology/NeuroKit
[2] github.com/Aura-healthcare/ecg_qc

that overlap with Neurokit2 and shares the issue of having no evaluation results on open source datasets.

In this work, we develop and analyze an open-source, reproducible, python implementation of many common SQA methods, including signal quality indices for analysis of electrocardiogram (ECG), plethysmography (Pleth), and other vital sign timeseries data. Additionally, we validate our results on both a closed-source bleeding dataset from the University of Pittsburgh Medical Center (UPMC) [70] as well as publicly available datasets such as the MIT-BIH Arrhythmia dataset [67] (See Figure 1.2).

Our main task is to analyze signal quality from recorded Pig Lab Experiments with University of Pittsburgh School of Medicine. The data consists of 25 pigs are anesthetized, bled at 10 mL/min, and resuscitated when hemorrhagic shock is observed (See Figure 1.1).

Figure 1.1: Diagram of the data colletion process for the Pig Data. First, the Config stage consists of calibration of equipment. Next, the baseline stage consists of monitoring stable pig baseline data to have a comparison for later use. Next, The Bleed stage is when the pig is bled at 10 mL/min until hemorrhagic shock threshold is reached. Finally, The Wait / resuscitate stage attempts to resuscitate the pig by adding fluid and medicine until the pig is stable.

Although 25 patients is current, the number is consistently increasing due to ongoing experiments. Vital signs, as well as other features, are recorded over the course of the hemorrhage process. This includes 9 Signals: Plethysmography (Pleth), Electrocardiogram (ECG), Cardiac Output (CO), Venous Oxygen Saturation (SvO2), Pulsoxymetrically Measured Oxygen Saturation (SpO2), Pulmonary Arterial Pressure (PAP), Central Venous Blood Pressure (CVP), Arterial Blood Pressure (ART), and finally, Air Pressure in Lungs (AirPr). Overall, this data collection yields 3 to 7 hrs of data at 250hz.

After discussing with data collection staff as well as clinicians, there are 2 primary issues

1. Issue 1: Detection of anomalous signals. Solution: Outlier detection via Signal

Quality Indices (SQIs)

2. Issue 2: Noisy ECG Signal. Solution: ECG denoising techniques, e.g. wavelet denoising

For outlier detection on signals, Signal Quality Indices (SQIs) are a natural featurization. To demonstrate the power of these SQIs, we produce results on 2 case studies. Firstly, we reproduce an old Physionet 2011 ECG Quality Classification Challenge [94] with 8 different methods, codify lessons learned, and show that a combination of all methods produce the best results, as opposed to any one method. Secondly, we demonstrate the ability of SQIs to improve Real vs Artifactual alert detection [34].

Next, we perform outlier detection using the SQIs introduced previously. We empirically find the best outlier detection method out of 4 methods on the Physionet dataset and show results on the pig dataset.

Finally, we will go over ECG denoising. We again, empirically find the best method out of 3 traditional methods on the open source MIT-BIH-Arrhythmia [67] and its corresponding Noise Stress Test Dataset [68], and show results on the pig dataset.

We will also release our code, which will be completely built on top of open source python packages and complete to reproduce baseline results out of the box (See Figure 1.2).

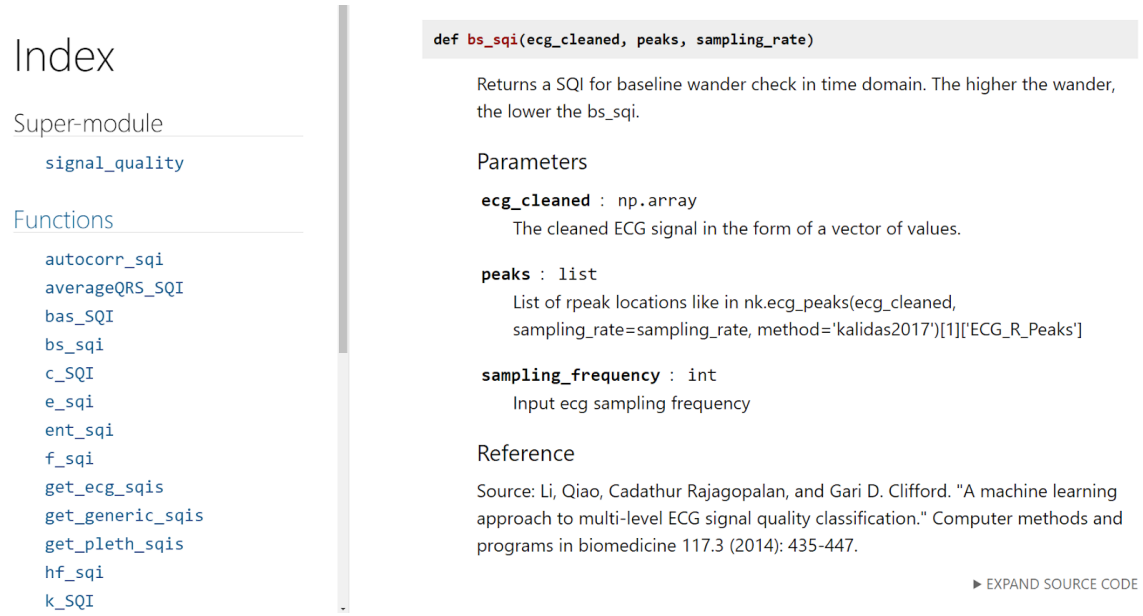


Figure 1.2: github.com/chufangao/signal_quality is an open-source, reproducible, python implementation of many common Signal Quality Analysis methods, including signal quality indices for analysis of electrocardiogram (ECG), plethysmography (Pleth), and more. In addition, example notebooks containing code for signal quality classification, outlier detection, and ECG Signal Denoising is also provided. Our methods are evaluated on publicly available datasets like the MIT-BIH Arrhythmia dataset and the Physionet 2011 ECG Quality Classification Challenge.

1.1 Examples of Noise

Common artifacts in vital sign recording include the following. [12, 45, 91]

1. **Baseline Wander and Abrupt Drift**. This is essentially a baseline noise that occurs while collecting ECG data, that may be caused by respiration (sedated or not), electrolyte properties, skin impedance, and other sources of random noise. See Figure 1.3.
2. **Electrode motion artifact**. This type of noise is caused by intermittent mechanical noise and may mimic the appearance of beats. It is generally considered the most troublesome, since it cannot be removed easily by simple filters. See Figure 1.4.
3. **Sudden Body Movement**. This includes artifacts induced by electrical

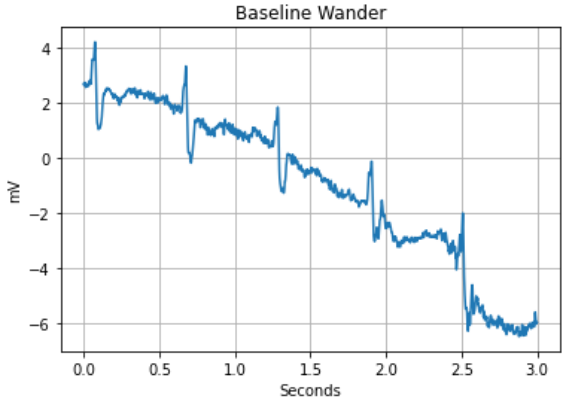


Figure 1.3: Baseline Wander Example in ECG

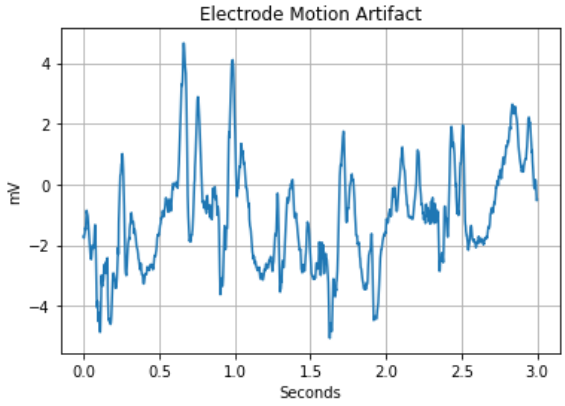


Figure 1.4: Electrode motion artifact Example in ECG

activity of muscles during periods of contraction, which is usually an issue for not-sedated subjects. However, induced movement of a sedated subject, such as via manual blood draws, also causes many artifact in the vital sign signals. See Figure 1.5.

4. **Measurement Equipment Failure**. This area includes areas such as mistakenly reversing the ECG leads on the subject body, or simply, the measurement equipment not being calibrated. E.g. Flatlines, wrong amplitude, losing one electrode, etc. See Figure 1.6. Additionally, during experiments, if a resuscitation pump fails, the ECG would not match a physician's intuition of a recovery trajectory. Therefore, extra care should be taken to interpret data from affected subjects accordingly.

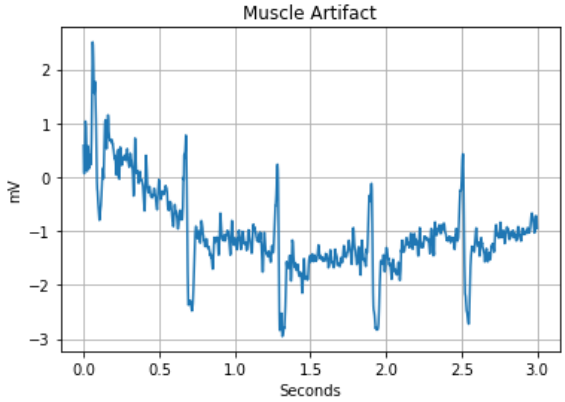


Figure 1.5: Muscle Artifact Example in ECG

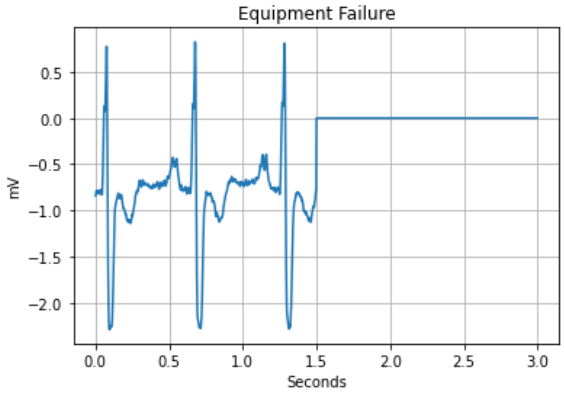


Figure 1.6: Equipment Failure Example in ECG

5. **Power-Line Interference (50-60 hz Signal Interference)** It is common that measurement equipment absorbs a constant 50-60 hz noise cause by inductive and capacitive couplings of ubiquitous power lines in the ECG signal acquisition circuitry. However, due to its ubiquitous nature, it is usually easily removed by notch filters.

Chapter 2

Literature Review

2.1 ECG Signal Quality

Signal Quality Index (SQI / SQA) Methods

Beat Detection, Inter-channel Beats, Kurtosis, and Spectral SQI: Li et al. [52] was one of the first to apply machine learning to address ECG signal quality. Although the main focus of that work is robust heart rate estimation, it introduces some common ECG SQIS that many future works use. The dataset used was the MIMIC-II [87], where real ECG noise from the MIT-BIH Noise Stress Test Database [68] was added to clean ECG signals. ECG signal quality was calculated via 4 methods.

First, two beat detection algorithms, wqrs [106] and ep_limited [28] are compared.

$$bSQI = \frac{\text{Number of beats detected by both}}{\text{Number of beats detected in total}}$$

bSQI is calculated for each heartbeat with a window of 10 seconds.

Second, an inter-channel signal quality metric iSQI is calculated as

$$iSQI = \max_{\text{over 10 seconds}} \frac{\text{Number of matched beats}}{\text{Number of all detected beats}}$$

Third, the kurtosis of the signal is calculated

$$kSQI = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right]^4$$

where $\hat{\mu}, \hat{\sigma}$ are the empirical mean and standard deviation of the discrete signal x of length M (i.e. signal can be stored in an array of length M).

Fourth, the spectral distribution of ECG is taken into account. The spectral distribution ration

$$SDR(\text{kth beat}) = \frac{\int_{f=5hz}^{f=14hz} P(\text{10s window around kth beat})df}{\int_{f=5hz}^{f=50hz} P(\text{10s window around kth beat})df}$$

Finally, pSQI was defined $pSQI = \begin{cases} 1 \text{ if } .5 \leq SDR \leq .8 \\ 0, \text{ otherwise} \end{cases}$

Finally, the SQIs are used along with Kalman Filtering to show that using SQIs significantly improved the heart rate estimation.

Skewness, Flat Line, and Baseline Power SQI: Clifford et al. [15] use the same 4 SQIs as Li et al. [52] above, but also add 3 more SQIs from [13]. The authors were able to achieve the second-highest score [53] in the PhysioNet/Computing in Cardiology Challenge 2011 event 1 [94]. To obtain balanced samples of all 3 signal quality groups (acceptable, intermediate, unacceptable), noise was added to clean ECG signals from noise stress test database [68].

The authors added a skewness SQI

$$sSQI = \frac{1}{M}\sum_{i=1}^{M}\left[\frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i}\right]^3$$

where $\hat{\mu}_i, \hat{\sigma}_i$ are the empirical mean and standard deviation of x_i.

Additionally, a flatness SQI: fSQI was added that quantified how long the signal was a flat line.

Finally, basSQI, or the the relative power in the baseline, was added, which was

calculated as the following:

$$basSQI = 1 - \frac{\int_{f=0hz}^{f=1hz} P(\text{ecg\_window})df}{\int_{f=0hz}^{f=40hz} P(\text{ecg\_window})df}$$

Finally, the authors compared 2 ML models–Support Vector Machine (SVM) and Multilayer Perceptron (MLP)–on the ECG signal quality classification task. Although both achived similar accuracy, the SVM generally performed slightly better, achieving an accuracy of 98% on the training data and 97% on the test data.

Baseline Wander, Energy, Complexity, Entropy, and other SQIs: Li et al. [55] developed a five-level, machine learning based, signal quality classification algorithm. In total, 13 signal quality metrics were derived from segments of ECG waveforms and classified via support vector machine. Training was done on a simulated dataset and validation was done on the MIT-BIH arrhythmia database [67]. The PCinC Challenge dataset [94] was re-annotated, adding varying levels of noise (varying signal to noise ratio levels) to the clean signals from the MIT-BIH Noise Stress Test Database (NSTDB) [68]. The validation dataset was also re-annotated. After QRS detection, 6 more SQIs were added to the SQIs used by Clifford et al. [15] The following list of SQIs were extracted:

Like previous work, the fraction of beats detected by 2 different QRS detectors (bSQI), the third moment or skewness of the ECG signal (sSQI), the fourth moment or kurtosis of the ECG signal (kSQI), the relative power in the QRS complex (pSQI), the relative power in the baseline (basSQI) was extracted.

Additionally, a baseline wander check in time domain (bsSQI) was extracted.

$$bsSQI = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\text{peak-to-peak amplitude of QRS}}{\text{peak-to-peak amplitude of filtered baseline}}\right)$$

The relative energy in the QRS complex was calculated:

$$eSQI = \frac{\sum_i E(QRS_i)}{E(window)}$$

where $E(x) = \sum_i^{|x|} x_i^2$ is the energy of the time series signal, and QRS_i is a detected

QRS signal segmented by $[R-.7s, R+.8s]$ and *window* is the time window of analysis.

The relative amplitude of high frequency noise is calculated as the following:

$$hfSQI = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{peak to peak amplitude of QRS}}{H_i}$$

where the ECG signal was filtered through an integer filter $y(j) = x(j) - 2x(j-1) + x(j-2), s(j) = |y(j)| + |y(j-1)| + \cdots + |y(j-5)|, H_i = mean_{[R-.28s,R-.05s]}(s(j))$

The signal purity[73] (Hjorth Descriptor [25]) of ECG is calculated:

$$purSQI = \frac{\bar{\omega_2}(k)^2}{\bar{\omega_0}(k)\bar{\omega_4}(k)}$$

where $\bar{\omega_n}$ is the n-th order spectral moment $\omega = \int_{-\pi}^{\pi}\omega^n P(e^{j\omega})d\omega$, with angular frequency $\omega = 2\pi f$

The relative standard deviation (STD) of QRS complex is calculated:

$$rsdSQI = \frac{1}{N}\sum_{i=1}^{N}\frac{\sigma r_i}{2\sigma a_i}$$

where $\sigma r_i = STD_{[R-.07s,R+.08s]}(QRS_i)$ is the standard deviation around a small window of the QRS signal and $\sigma r_i = STD_{[R-.2s,R+.2s]}(QRS_i)$ is the standard deviation around a large window of the QRS signal.

The sample entropy[85] of the ECG waveform is calculated:

$$entSQI = En(m, r, N) = -\ln\left[\frac{A^m(r)}{B^m(r)}\right]$$

where N is the length of the ECG signal ((i.e. signal can be stored in an array of length N)), and $En(m, r, N)$ as per SampEn in Richman et al. [85].

The high frequency mask of the ECG waveform is calculated

$$hfMSQI = \frac{\sum_i p_i}{p_a}$$

where p_i is the instantaneous power estimate of the ith QRS and p_a is the instantaneous power estimate of the full window

Finally, a periodic component analysis (PiCA) periodicity measure [89] of the ECG waveform is calculated

$$PiCASQI = |\frac{C_x(\tau_t)}{C_x(0)}| = |\frac{E_t\{x(t+\tau_t)x(t)\}}{E_t\{x(t)^2\}}|$$

where $C_x(\tau_t) = x(t+\tau_t)x(t)\}$ is the covariance of the signal x a lag of τ_t on x. The lag is applied such that x is phase-aligned with the previous beat.

2 metrics was use to measure performance: classification accuracy, and a single class overlap accuracy, which assumes that an individual type classified into an adjacent class is acceptable. 5-fold cross validation with an SVM yielded a classification accuracy of 88.07 $\pm$ 0.32% and a single class overlap accuracy of 99.34 $\pm$ 0.07% on MITDB. The test set yielded an a classification accuracy of 57.26% and an a single class overlap accuracy of 94.23%.

PCA SQI: Behar et al. [5] also used ECG SQIs and featurizations to reduce false alarms (FA) in ICU monitors via SVM. The PICC dataset [94] was used to train a quality assessment model. The MIT-BIH arrhythmia dataset [67] was used to test model performance on arrhythmic records as well as on records of a different modality, and the MIMIC-II dataset [88] was used for discovering associations between quality and FA in ICUs. In this work, only binary classification of ECG signal quality was considered.

7 SQIs were calculated on 10 second intervals of each ECG lead: the fourth moment (kurtosis) of the signal kSQI, the third moment (skewness) of the signal sSQI, the relative power in the QRS complex pSQI, the relative power in the baseline basSQI, the fraction and ratio of beats detected by 2 different QRS detectors bSQI.

A new SQI is added, which is:

$$pcaSQI = \frac{\lambda_1 + \lambda_2 + \cdots \lambda_5}{\sum_i \lambda_i}$$

where λ_i is the ith eigenvalue associated with the ith principal components obtained by principal component analysis of all detected QRS windows.

Results from SVM with a Gaussian Kernel showed classification accuracies of 89% overall (on all arrhythmia types), although performance varied greatly depending

on the type of rhythm (arrhythmia type), indicating that SQIs should be rhythm specific and that classifiers should be trained for each rhythm call independently.

QRS Template Matching: Orphanidou et al. proposed a Signal Quality Indice (SQI) for Electrocardiogram (ECG) and Photoplethysmogram (PPG) [74] to measure Heart Rate calculation ability, with a binary output of "good" indicating that a reliable HR can be derived and "bad" indicating otherwise. First, a sanity check of the HR being larger than 40 and smaller than 180 beats per minute is done. Secondly, a check to make sure that there should be no more than a 3 second gap between each beat was completed. Finally, another sanity check is made such that, in each 10 second sample of ECG, the ratio of the largest and the smallest beat-to-beat interval should be less than 10%. After these 3 checks were performed, individual QRS complexes and PPG pulse waves were extracted and averaged to create a template. Finally, correlation coefficients were used to calculate the average correlation between each window and its template. The average correlation was then thresholded with a cross validated cutoff value to determine "goodness" of signal. The authors found that the optimum threshold for the average correlation coefficient was 0.66 for the ECG SQI and 0.86 for the PPG SQI. This corresponded to a sensitivity of 94% and specificity of 97% for the ECG and a sensitivity of 91% and specificity of 95% for the PPG SQI [74].

Rpeak Detection and Power Spectrum SQIs: Zhao et al. [105] extracted 6 SQIs from the ECG signal: the matching degree of R peak detection, power spectrum distribution of QRS wave, variability in the R-R interval, kurtosis, skewness, and baseline relative power. The features were used in a simple heuristic fusion as well as a fuzzy comprehensive evaluation. The 4 best SQIs combination based on simple combinations and thresholding–qSQI, pSQI, kSQI, basSQ–were used in a rule-based classification system. The system assigned classifications based off of simple manual thresholds for each SQI. An open source implementation of both methods is available in [61].

Neurokit SQIs: The Python packages Neurokit and Neurokit2 [61] has 3 versions of ECG Signal Quality metrics. The older Neurokit has a method[1] that computes the predicted probability of given ECG signal coming from the primary lead. This prediction is made from a standard Sklearn MultiLayerPerceptron (MLP) Classifier [77] trained on the PTB-Diagnostic dataset [6] available from PhysioNet [26]. The training setup is as follows: First, extract all the ECG signals from the healthy participants, obtaining 15 recording leads per subject. Second, for each lead, downsample the signal from 600 to 200 datapoints and extract all cardiac cycles. Third, the MLP model is fit on 2/3 of the dataset (134392 cardiac cycles) to predict the lead, with evaluation on the remaining 1/3. This yields an accuracy of 0.91 and a precision of 0.91 [61].

The newer Neurokit2 calculates a correlation of the QRS waveform to the average QRS waveform (averageqrs_sqi) – a template matching method [2]. It calculates a continuous index of quality of the ECG signal by interpolating the distance of each QRS segment from the average QRS segment present in the data. The interpolation is performed via a 2nd-order spline. An averageqrs_sqi of 1 corresponds to heartbeats that are the closest to the average sample and an averageqrs_sqi of 0 corresponds to the most distant heartbeat from that average sample. The authors note that 1 does not necessarily means "good": if the average of the QRS windows are bad, than being close to the average will not be optimal. Finally, Neurokit2 also supports the method specified by Zhao et al. [105].

2.2 ECG Denoising

Wavelet Denoising: Wavelets are an established method of denoising signals by discrete wavelet decomposition (DWT), thresholding the decomposed coefficients, and inverse wavelet decomposition (IDWT) [8, 16, 99].

The discrete wavelet transform works as follows [78]. The input signal x is calculated by passing it through a series of filters. First x is convolved with a low-pass

[1] https://github.com/neuropsychology/NeuroKit.py/blob/master/neurokit/bio/bio_ecg.py#L328

[2] https://github.com/neuropsychology/NeuroKit/blob/master/neurokit2/ecg/ecg_quality.py#L16

filter with impulse response g. x is also convolved with high-pass filter h.

$$y_{low}[n] = (x * g)[n] = \sum_{k=\infty}^{\infty} x[k]g[n-k]$$

$$y_{high}[n] = (x * h)[n] = \sum_{k=\infty}^{\infty} x[k]h[n-k]$$

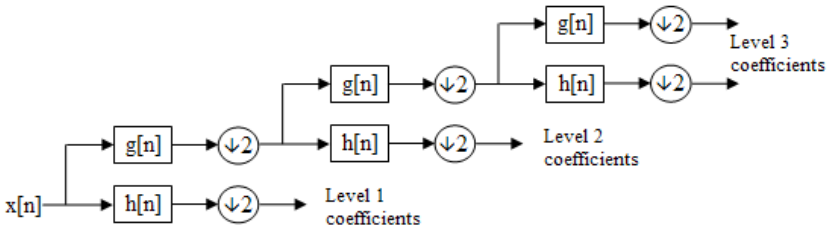


Figure 2.1: Cascading representation of DWT [44]. At each level, the signal is decomposed into low and high frequencies. The input signal must be a multiple of 2^n where n is the number of levels. E.g. a signal with 32 samples, frequency range 0 to f_n and 3 levels of decomposition, 4 output scales are produced.

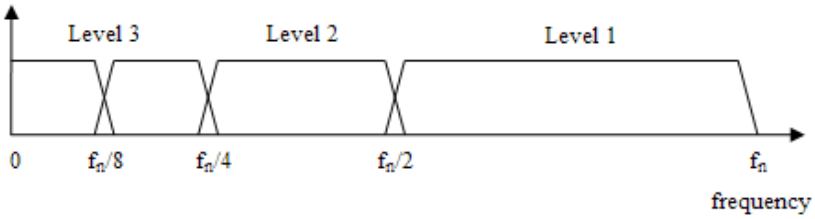


Figure 2.2: Frequency domain representation of DWT [43]

y_{low} is known as the approximation coefficients, and y_{high} is known as the detail coefficients. Now, since half of the frequencies of the signal have been separated, we can down-sample the signal by a factor of 2 with loss according to Nyquist's Rule. We downsample y_{low} and process it again by passing it through a new low-pass filter g and high-pass filter h. This process is then repeated, as shown in Figure 2.1 and Figure 2.2. Finally, g and h have a special relationship. They are known as filter coefficients and fit the properties of compactness and orthogonality.

$$g_k = (-1)^k) h_{n-k-1}, k \in \{0, \ldots, n-1\}$$

For ECG denoising, it is common to use the wavelet Daubechies-4 (db4)[3] , the most compact of which has $h = [\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{(3+\sqrt{3})}{4\sqrt{2}}, \frac{(3-\sqrt{3})}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}]$

For example, Singh et al. [96] presents a selection procedure of mother wavelet basis functions applied for denoising of the ECG signal in wavelet domain while retaining the signal peaks close to their full amplitude. The obtained wavelet based denoised ECG signals retain the most important information contained in the original ECG signal–the ECG peaks. Similarly, Alfaouri et al. [3] presents an approach based on the threshold value of ECG signal determination using Wavelet Transform coefficients that performs better than Donoho's method [16] in terms of signal to noise ratio (SNR) due to better thresholding functions on the decomposed coefficients.

However, it is important to point out that many of these methods are testing on primarily white noise, such as gaussian noise, and that more complicated sources of noise are lacking.

Empirical Mode Decomposition (EMD) Empirical Mode Decomposition is a data-driven decomposition method that does not require any basis, unlike DWT [21, 86]. The goal of EMD is to decompose the signal into a sum of Intrinsic Mode Functions (IMFs), which are like DWT coefficients. Ideally, this decomposes a signal into physically meaningful components. When all IMFs are summed up, the original signal is returned. An IMF is defined as 1) function with equal number of extrema and zero crossings (or at most differed by one) and 2) The envelopes connecting the local minima and local maxima of an IMF have a local average of zero [97, 103].

The process of extracting IMFs from a given input signal is called the "sifting process". Given a signal x, we first extract all the local minima and local maxima. Then, let the upper envelop e_u be a cubic spline curve fitted on all the local maxima. Similarly, let the lower envelop e_l be a cubic spline curve fitted on all the local minima. Let the average envelope of the upper and lower envelopes be $m_1 = (e_u + e_l)/2$. The

[3] http://wavelets.pybytes.com/wavelet/db4/

first proto-IMF is the original signal minus the average envelope

$$h_1 = x - m_1$$

The sifting process is iteratively performed again on h_1 as the input signal until we obtain proto-IMF h_k that satisfies the stopping criterion. Then, we let the $c_1 = h_k$ be the first IMF. A commonly used stopping criterion is the Sum of Difference (SD)

$$SD = \sum_{t=0}^{T} \frac{|h_{k-1}(t) - h_k(t)|^2}{h_{k-1}^2}$$

When the SD is smaller than a threshold, the first IMF c_1 is obtained, which is written as

$$x - c_1 = r_1$$

Note that the residue r_1 still contains useful information. We can therefore treat the r_1 as a new signal and reapply the sifting procedure with the stopping criterion to obtain:

$$r_{i-1} - c_i = r_i \quad i = 1, \ldots, N$$

The whole procedure terminates when the residue r_N is either a constant, a monotonic slope, or a function with only one extremum. Since each IMF is simply a linear relationship with the other IMFs and the original signal, we can write:

$$x = \sum_{n=1}^{N} c_n + r_N$$

Weng et al. [103] proposed am ECG denoising method based on Empirical Mode Decomposition, which is able to remove high frequency noise with minimum signal distortion. This is accomplished by adding an additional step so that QRS signals are preserved, since otherwise, their amplitudes would be decreased. Evaluation on the MIT-BIH database show that the proposed method is successful in denoising ECG with added Gaussian noise.

Fully Convolutional Denoising Autoencoders: Recently, Neural Networks for the use of ECG denoising have been making great strides in terms of both accuracy and speed [79]. Chiang et al [11] introduced a wavelet denoising method using fully convolutional denoising autoencoders (DAE) for ECG. The proposed FCN-based DAE can also perform compression due to the encoder part of the architecture. Evaluation was performed on the MIT-BIH Arrhythmia database [67] with added noise signals from the MIT-BIH Noise Stress Test database [68]. Convolutional Autoencoders (CNN AEs) use 1D convolutions instead of recurrent neural networks (RNNs), which make them faster to train and more accurate. Results conducted on noisy ECG signals of different levels of noise show that the FCN achieves better performance compared to deep fully connected neural networks and convolutional neural network-based (without autoencoder) denoising models with lower RMSE and higher Signal to noise ratio (SNR) improvement.

Autoencoders (AE) denoise signals by data compression and reconstruction. First, a neural network is used to embed signal into a lower dimensional embedding. Then, a NN is used to reconstruct the signal from the embedding. The model is trained on noisy data as input with a mean squared error loss of the output and the clean data. This has the practical effect of training the model to learn the semantics of the input while ignoring the noise. Figure 2.3 shows this process, along with an example model architecture for the encoder and decoder. Note that the input to the encoder portion of the autoencoder may be a processed input. For example, [33] showed that using the ECG decomposed into coefficients via discrete wavelet transform as features to a Long Short-Term Memory (LSTM) Autoencoder was highly useful for arrhythmia classification. Because of this, we employ a similar method for denoising. In addition to the wavelet coefficients, we also pass in the original signal as input, with the inuition that the more information that the AE has, the better it can potentially perform.

2.3 Open Source Datasets

There are only a few common, open, publicly available, and well-documented datasets for evaluating ECG signal quality. The following is a list of some of the most common datasets.

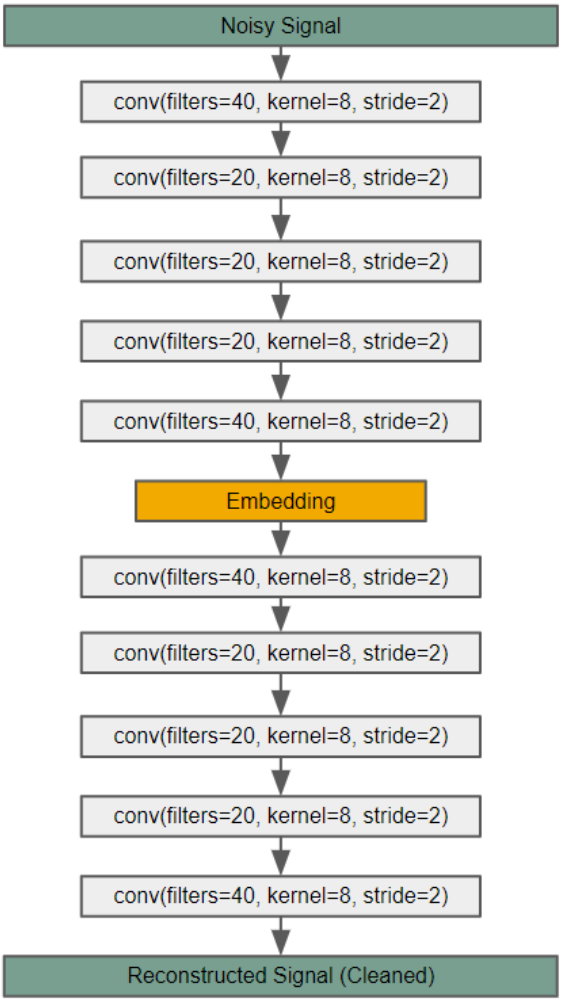


Figure 2.3: Example Convolutional Autoencoder Architecture. The input is noisy data and the output is the cleaned data.

MIT-BIH Arrhythmia Database [62, 66, 67] This is an open access dataset available on PhysioNet [26] collected and maintained by the Beth Israel Deaconess Medical Center and MIT between 1975 and 1979. It contains contains 48 half-hour two-channel ambulatory ECG recordings, from 47 subjects. 23 recordings were chosen at random from a mixed population of inpatients (about 60%) and outpatients (about 40%) at Boston's Beth Israel Hospital. 25 more recordings were selected to include less common, but clinically significant, arrhythmias that would not be well represented in a random sample. The recordings were digitized at 360 hz per channel with 11-bit

resolution over a 10 mV range. Two or more cardiologists independently annotated almost every QRS signal. Disagreements were resolved to obtain the computer-readable reference annotations for each beat (approximately 110,000 annotations in all) included with the database.

Since this is an arrhythmia database and not a signal quality database, it is common to the signals such that any beat that is not labelled as "Normal" is noisy.

MIT-BIH Noise Stress Test Database (NSTDB) [68] This is a database also available on PhysioNet [26],includes 12 half-hour ECG recordings and 3 half-hour recordings of noise typical in ambulatory ECG recordings. The noise recordings were made using physically active volunteers and standard ECG recorders, leads, and electrodes; the electrodes were placed on the limbs in positions in which the subjects' ECGs were not visible.

The three noise records were assembled from the recordings by ECG signals that contained predominantly baseline wander (bw), muscle (EMG) artifact (ma), and electrode motion artifact (em) noise. Baseline wander noise is usually low-frequency noise caused by motion. Muscle EMG artifact is usually electrical noise generated by muscle activity near the electrode. Electrode motion artifact is intermittent mechanical noise, and is generally considered the most troublesome, since it may mimic the appearance of beats and cannot be removed easily by simple filters.

The noisy ECG recordings were created by the script nstdbgen, a script provided in the database to manually add noise to clean signals. Using two clean recordings (118 and 119) from the MIT-BIH Arrhythmia Database, calibrated levels of noise from record 'em' were added. Noise was added beginning after the first 5 minutes of each record, during two-minute segments alternating with two-minute clean segments. Since the original ECG recordings are clean, the correct beat annotations are known even when the noise makes the recordings visually unreadable. The reference annotations for these records are simply copies of those for the original clean ECGs.

The PhysioNet/Computing in Cardiology Challenge 2011 dataset: Also known as the PICC dataset, this is another common dataset [94] found on PhysioNet [26]. The dataset includes ten-second recordings of twelve-lead ECGs, consisting of standard 12-lead ECG recordings (leads I, II, II, aVR, aVL,aVF, V1, V2, V3, V4, V5,

and V6) with full diagnostic bandwidth (0.05 through 100 Hz). Each lead is recorded simultaneously for a minimum of 10 seconds (and generally only a maximum of 10 seconds) at 500 Hz with 16-bit resolution. The annotation was conducted by nurses, technicians, and volunteers with varying amounts of training.

ECGs collected for the challenge were reviewed by a group of annotators with varying amounts of expertise in ECG analysis, ranging from volunteers to experts, in blinded fashion for grading and interpretation. 3 to 18 annotators independently examined each ECG, assigning it a letter grade (A (0.95): excellent, B (0.85): good, C (0.75): adequate, D (0.60): poor, or F (0): unacceptable) for signal quality. The average grade was calculated in for each record, and fell into 1 of 3 groups:

Group 1 (acceptable): If average grade ≥ 0.70, and at most one grade is F. Group 2 (indeterminate): If average grade ≥ 0.70 or more, but two or more grades were F. Group 3 (unacceptable): If average grade is $<$ than 0.70.

Approximately 70% of the collected records were assigned to group 1, 30% to group 3, and fewer than 1% to group 2, reflecting a high degree of agreement among the annotators. For the purposes of this experiments, we use the annotations from this dataset that are explicitly annotated ACCEPTABLE vs UNACCEPTABLE.

2.4 Signal Quality in Other Waveforms

Although most work in the signal quality space is specifically regarding ECG, many of the metrics can be applied in the general case. For general signal quality, we use time series featurizations like means, ranges, standard deviations, sample entropy, and more. Furthermore, we adapt some features inspired by ECG SQIs, like percent of flat line in a signal, complexity measures like Hjorth Descriptor [25]. A full list of the adapted SQIs for both ECG and other signals, please see the file `sqis.py` in the released python package.

For example, Plethysmography (PPG) signal quality has many of the same metrics as ECG. Elgendi et al. [18] proposed eight SQIs to be tested and evaluated on their abilities to quantify "goodness" of PPG. The dataset consists of heat stress PPG data collected as part of a National Critical Care and Trauma Response Centre (NCCTRC) project [7]. This dataset has a total of 160 signals which are then further processed to 106 PPG recordings at 60 s each. The SQIs consist of: 1. Perfusion, also

known as the difference of the amount of light absorbed through the pulse of when light is transmitted through the measurement area. $Perfusion = \frac{y_{max} - y_{min}}{|\bar{x}|} \times 100$. 2. Skewness, a measure of the symmetry (or the lack of it) of a probability distribution. 3. Kurtosis 4. Entropy 5. Zero crossing rate 6. Signal-to-noise ratio 7. Matching of multiple systolic wave detection algorithms 8. Relative power.

2.5 Outlier Detection

Let us consider 4 popular Outlier Detection methods

K-Nearest Neighbor This is a classic method of detecting outliers that is based on the distance of a point from its kth nearest neighbor. Although there are numerous ways to calculate a kth nearest neighbor, the details will be omitted here. However, to aid in efficiency, it is possible to rank each individual point on its distance to its kth nearest neighbor and declare the top n points in this ranking to be outliers. Since anomalies should be far from all neighbors, they should have a higher distance than normal points. Additionally, it is appropriate to think of the outlier score as the distance to a sample's kth nearest neighbor. [4, 82]

Isolation Forest The Isolation Forest algorithm seeks to separate anomalous observations from other observations by computer the average number of random splits needed for an observation to reach a leaf node.

First, an isolation forest is an ensemble of decision trees. The trees are constructed via randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. Then, selecting another feature, selecting another value to split on, and so on. Random partitioning produces noticeably shorter paths for anomalies, as they should be outliers in some way, and therefore easier to split out. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.

Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node. Thus the outlier score is the average number of splittings required for a sample to reach a leaf (over the forest) [59, 60].

One-Class Support Vector Machine A one-class Support Vector Machine (SVM) can be thought of as a model that is fitted to the data, where anomalous data is not well predicted by it. A standard SVM with 2 classes fits the data using a hyperplane with the largest possible margin between data of both classes. Although One-Class SVM is similar, instead of using a hyperplane to separate two classes of instances, it project points to a higher dimensional space and uses a minimum hypersphere to fit all the data points in the feature space. Finally, the points lying outside the hypersphere are outliers, since anomalies will be far away from the other points [92].

AutoEncoder Using neural networks, AutoEncoders (AEs) map data into lower dimensional embedding space, and then reconstruct data from the respective embedding. AEs can be trained in an unsupervised fashion like PCA, and can be used to detect outliers via reconstruction error. Ideally, anomalies will have higher reconstruction error as the model is not trained on it as much as regular data [1].

Chapter 3

Methodology and Experiments

3.1 Outlier Detection Via Featurization

3.1.1 Case Study: Reproducing ECG Signal Quality Classification Results on the Physionet 2011 Challenge

The Physionet 2011 ECG signal quality classification challenge was created to quickly access the quality of ECG. There has been numerous publications regarding this challenge dataset, but much of the work that currently exists is not implemented in code. Rather, the methods are only discussed on paper. Our main goal is to answer the following question: Can we reproduce experiments from baseline work?

The Physionet 2011 Challenge Data is split into 3 annotated sets: Set-A, Set-B, and Set-C. Only Set-A is fully public, so cross validation is needed. Specifically, we use 5-fold group split cross-validation on the subjects, so we do not have any overlapping subjects in the train / test split. The dataset ends up having 998 12-lead ECG recordings, each at 10 or more seconds, recorded at 500 Hz. The task is still relatively simple, as we are using only the explicitly labelled clean and noisy data provided in the dataset files. This is a straightforward Binary classification: 0=not noisy, 1=noisy.

For our experiments, we consider 8 individual methods, as well as 1 combined method that include features from all 8 methods (See Figure 3.1). Additionally, as

Method	Signal Quality Indices (SQIs) Used
(1) Li et al. 2007	b_sqi, p_sqi, k_sqi
(2) Clifford et al. 2012	b_sqi, p_sqi, k_sqi, s_sqi, f_sqi, bas_sqi
(3) Behar et al. 2013	k_sqi, s_sqi, p_sqi, b_sqi
(4) Li et al. 2014	b_sqi, p_sqi, k_sqi, s_sqi, f_sqi, bas_sqi, bs_sqi, e_sqi, hf_sqi, pur_sqi, rsd_sqi, ent_sqi
(5) Geometric Features	Median, IQR, and Slope of HR. R-peak interval STD. Samp. entr., appx. entr., Relative power, and ratio of LF and HF bands
Misc Methods	(6) averageQRS_sqi, (7) zhao2018_sqi, (8) orphanidou2015_sqi
(9) Combined	All of the previously mentioned SQIs and features

Figure 3.1: The 8 methods are shown in the table on the right highlighted in red, with their corresponding SQI features.

some of the SQIs depend on the existence of multiple leads of ECG (i_SQI and pca_SQI), we run an experiment with only single-lead SQIs, that is, with all SQIs except for i_SQI and pca_SQI. In the real world (and also in the pig dataset), we may not have access to more than 1 ECG signal. The specifics are in the full list of the adapted SQIs for both ECG and other signals (please see the file `sqis.py` in the released python package). Let us use the features that each of the 8 papers use as input to a Random Forest classifier, to account for subtle differences in modeling.

The following list is a description of each SQI, in order of introduction:

- Li et al. 2007 [52]
  - b_SQI - ratio of agreed beats detected by 2 algorithms
  - i_SQI - ratio of agreed beats detected by 2 leads
  - k_SQI - kurtosis
  - p_SQI - Power Spectrum Distribution
- Clifford et al. 2012 [15]

  - SQIs listed in Li et al. 2007
  - s_sqi - skewness
  - f_sqi - percentage of flat signal
  - bas_sqi - measures Relative Power in the low frequency baseline
- Behar et al. 2013 [5]
  - SQIs listed in Li et al. 2007
  - s_sqi - skewness
  - pca_sqi - sum of first 5 eigenvalues of pca decomposition
- Li et al. 2014 [55]
  - SQIs listed in Li et al. 2007
  - bs_sqi - baseline wander check
  - e_sqi - energy of detected QRS waveforms
  - hf_sqi - relative amplitude of high frequency noise
  - pur_sqi - Hjorth complexity (compares the signal's similarity to a pure sine wave)
  - rsd_sqi - relative standard deviation of QRS
  - ent_sqi - sample entropy
- Orphanidou et al. 2015 [74]
  - orphanidou2015_sqi - An SQI that measures the correlation between QRS and the mean QRS
- Neurokit [61]
  - averageQRS_sqi - A Neurokit function, An SQI that measures euclidean distance between QRS and the mean QRS
- Zhao et al. 2018 [105]
  - zhao2018_sqi - A Neurokit function, uses pSQI, kSQI and basSQI
- Geometric ECG Features [10] - Median, IQR, and Slope of Heart Rate, Standard deviation of R-peak intervals, Sample entropy, approximate entropy, Absolute power, Relative power, and ratio of LF and HF bands

- Combined: All features previously mentioned

Results on Open Source Dataset

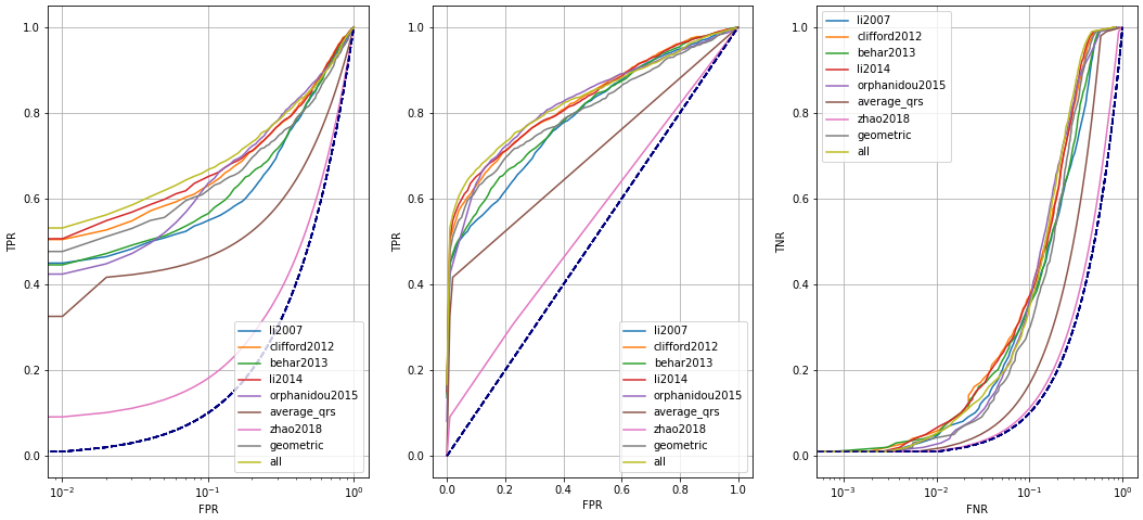


Figure 3.2: Results on subject-based 5-fold cross validation and Random Forest with *only* single-lead features

Method	AUC	Accuracy
li2007	0.793 ± 0.033	0.869 ± 0.013
clifford2012	0.829 ± 0.041	0.877 ± 0.014
behar2013	0.803 ± 0.035	0.869 ± 0.013
li2014	0.830 ± 0.041	0.882 ± 0.013
orphanidou2015	0.826 ± 0.028	0.863 ± 0.012
averageqrs	0.700 ± 0.025	0.858 ± 0.013
zhao2018	0.549 ± 0.013	0.792 ± 0.002
geometric	0.807 ± 0.028	0.875 ± 0.013
all	**0.835 ± 0.033**	**0.888 ± 0.012**

Table 3.1: Table of AUCs and Accuracies of 8 methods on subject-based 5-fold cross validation and Random Forest with *only* single-lead features

We see the results in Figure 3.2 and in Table 3.1. Additionally, we see that, compared to the original PhysioNet/Computing in Cardiology Challenge 2011 results in Table 3.2, the accuracy performance of the random forest ran on single-lead features

Participant	**Accuracy**
Xiaopeng Zhao	0.914
Benjamin Moody	0.896
Lars Johannesen	0.880
Philip Langley	0.868
Dieter Hayn	0.834
Václav Chudáček	0.833
Unofficial entries	
George Moody	0.894
Ikaro Silva	**0.802**

Table 3.2: The PhysioNet/Computing in Cardiology Challenge 2011 Official Results: Event 2 (open-source, open data set B), the most comparable event to our case. However, this is still not exactly a proper comparison, since we do not have access to data set B.

is comparable 3rd on the leaderboard, as shown in Table 3.1, suggesting that we are able to reproduce simple baselines. Furthermore, we see that neurokit's methods do not fare well under actual evaluation, especially zhao2018.

3.1.2 Case Study: Using SQIs to Improve Real vs Artifactual Alert Detection

High rates of false alarms for cardiorespiratory instability (CRI) in monitored patients cause alarm fatigue. Do our ECG SQIs help classification of artifacts in real-world human data to help prevent these false alarms?

Our dataset is a private dataset PPINNC – a set of time series data from Intensive Care Unit patients from University of Pittsburgh Medical Center. The PPINNC project was created to look at episodes of Cardio-Respiratory Insufficiency (CRI) with two primary objectives: 1. The first objective is to distinguish between real and artifactual CRI alerts. 2. The second objective is to predict when a patient is likely to have a CRI in the future.

There are 3 types of alerts in this dataset, corresponding to the following vital signs: SpO2, Heart Rate (HR), Respiratory Rate (RR). The vital sign timeseries are featurized into 6537 3-min windows, every 20 seconds. In the end, we end up with 255 features calculated from Pleth and ECG, including entropy, linear trends, heart

rate features, and more. This is a binary classification: Real Alert=1, Artifactual Alert=0, on the given features.

To maintain consistency with existing methods, we will use Random Forest, with patient level random splits (10x) for cross evaluation.

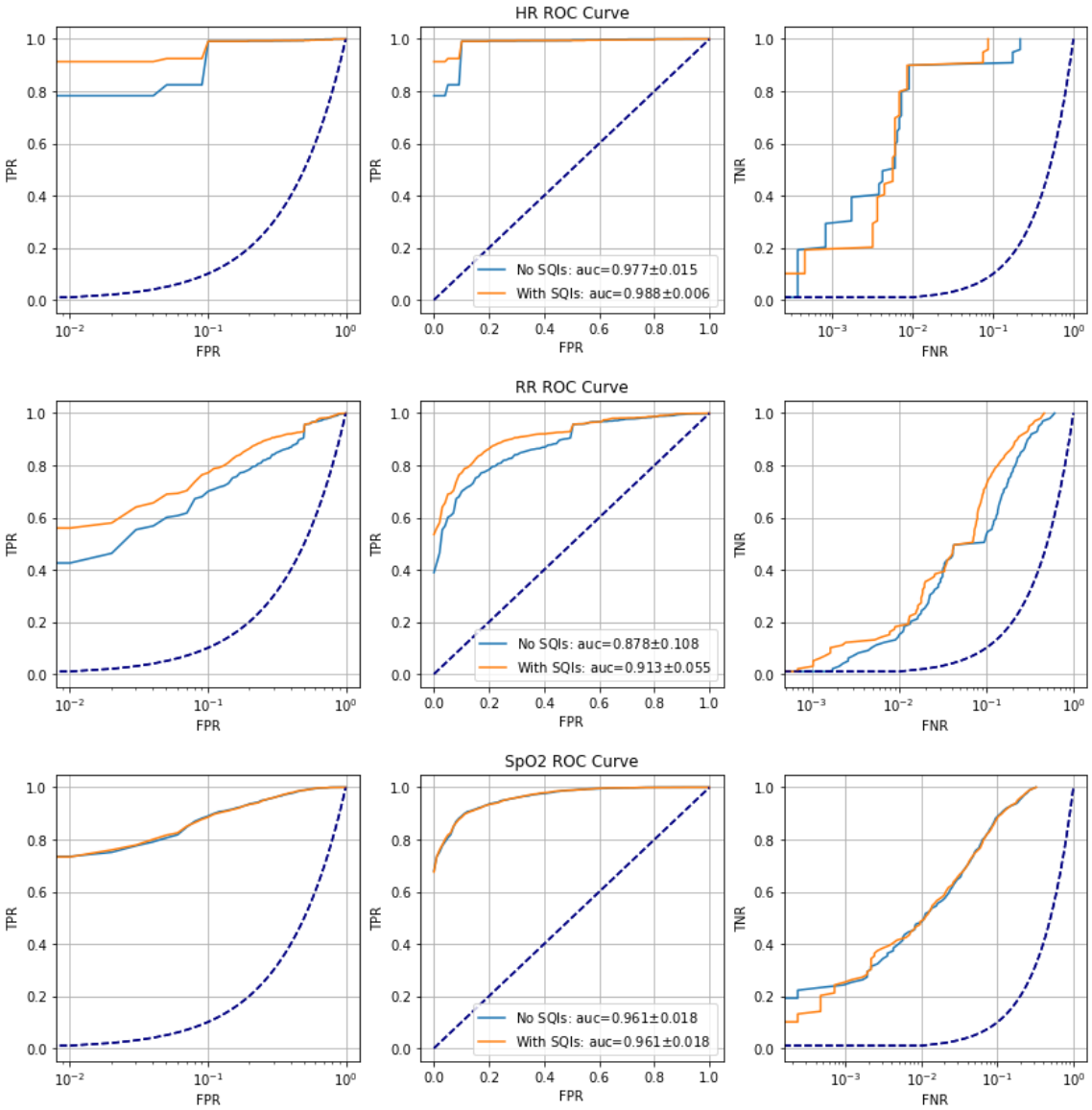


Figure 3.3: Plots of ROC Curves for all 3 alert types, with and without additional SQIs. For most alerts, we do improve AUC.

As seen in Figure 3.3, the SQIs that we added generally slightly improve AUC,

even with the extensive list of 255 precomputed features. Only the SpO2 ROC does not improve. RR alert classification improves the most in both true positives and true negatives. This shows that SQIs are encoding useful information, and are useful for classification of artifacts. Although this is a limited case study, further investigation should be done as to the type of SQIs that most significantly affact model performance.

3.1.3 Outlier Detection

Outlier Detection On PhysioNet Dataset

Let us consider 4 popular Outlier Detection methods: K-Nearest Neighbor, Isolation Forest, One-Class SVM, and AutoEncoder. First, we choose the best model based off of the PhysioNet/Computing in Cardiology Challenge 2011 dataset and 5-fold cross-validation.

For implementation of these algorithms, we follow the example in the open-source, anomaly detection Python package PyOD[1]. PyOD implements 30 anomaly detection algorithms, with evaluation results on 55 benchmark datasets. Specifically, each model is instantiated as in the example code[2], trained on the train split of the Physionet Dataset, and then evaluated on the test split. On the training set, each model is trained on a mix of clean and noisy signals, with the percentage of noisy signals in the dataset KNOWN. This is so that each model can find its own threshold for anomaly classification Finally, each model is tested to see if its predicted outliers match the noise classification in the test split.

We can use all the single-lead SQIs from the previous case study. To maintain fairness, we train all models in an unsupervised fashion. Secondly, once we have determined the best model, we can use it and apply it to the pig data as mentioned in the introduction.

From the results in Figure 3.4 and in Table 3.3 we see that Isolation Forest performs the best in all metrics. So, we will use Isolation Forest for the real-world pig data.

[1] pyod.readthedocs.io
[2] pyod.readthedocs.io/en/latest/example.html

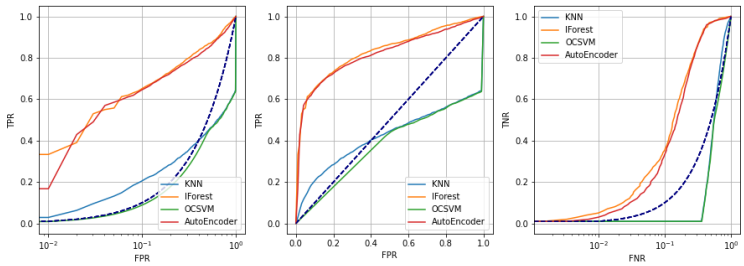


Figure 3.4: Plots of ROC Curves for all 4 different Outlier Detection algorithms on subject-based 5-fold cross validation

Method	Train AUC	Train Accuracy	Test AUC	Test Accuracy
KNN	0.421 ± 0.035	0.649 ± 0.012	0.396 ± 0.013	0.693 ± 0.003
IForest	**0.832 ± 0.024**	**0.843 ± 0.020**	**0.830 ± 0.006**	**0.844 ± 0.004**
OCSVM	0.380 ± 0.033	0.157 ± 0.015	0.492 ± 0.062	0.646 ± 0.015
AutoEncoder	0.819 ± 0.020	0.842 ± 0.019	0.819 ± 0.007	0.843 ± 0.003

Table 3.3: Results of 4 different OD algorithms on subject-based 5-fold cross validation.

Outlier Detection On Pig Dataset

However, there are a few things to consider in the real-world pig dataset. First, it is important to train outlier detectors on ”normal” data so that they are actually sensitive to outliers. Since the pig hemorrhage experiment only starts after a certain amount of time, most data before this baseline time should be discarded, as it is quite noisy. Furthermore, we should also not use the end of the recording, as it may also be quite noisy (See Figure 3.5).

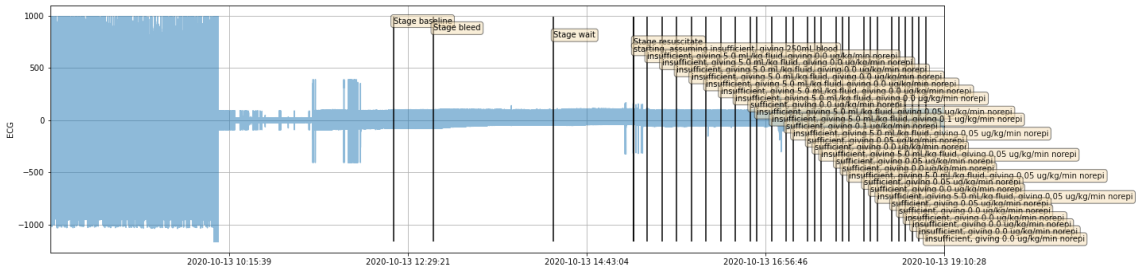


Figure 3.5: Example of Noise in the ECG Vital Sign Time Series Signal for the Pig Data. The raw signal is shown in blue, with annotations indicated by the black vertical lines.

For brevity, we will only discuss Pig08; the full list of results can be found in Appendix B. First off, we show the results of all the outliers in the form of a plot, scaled to the original time-series data.

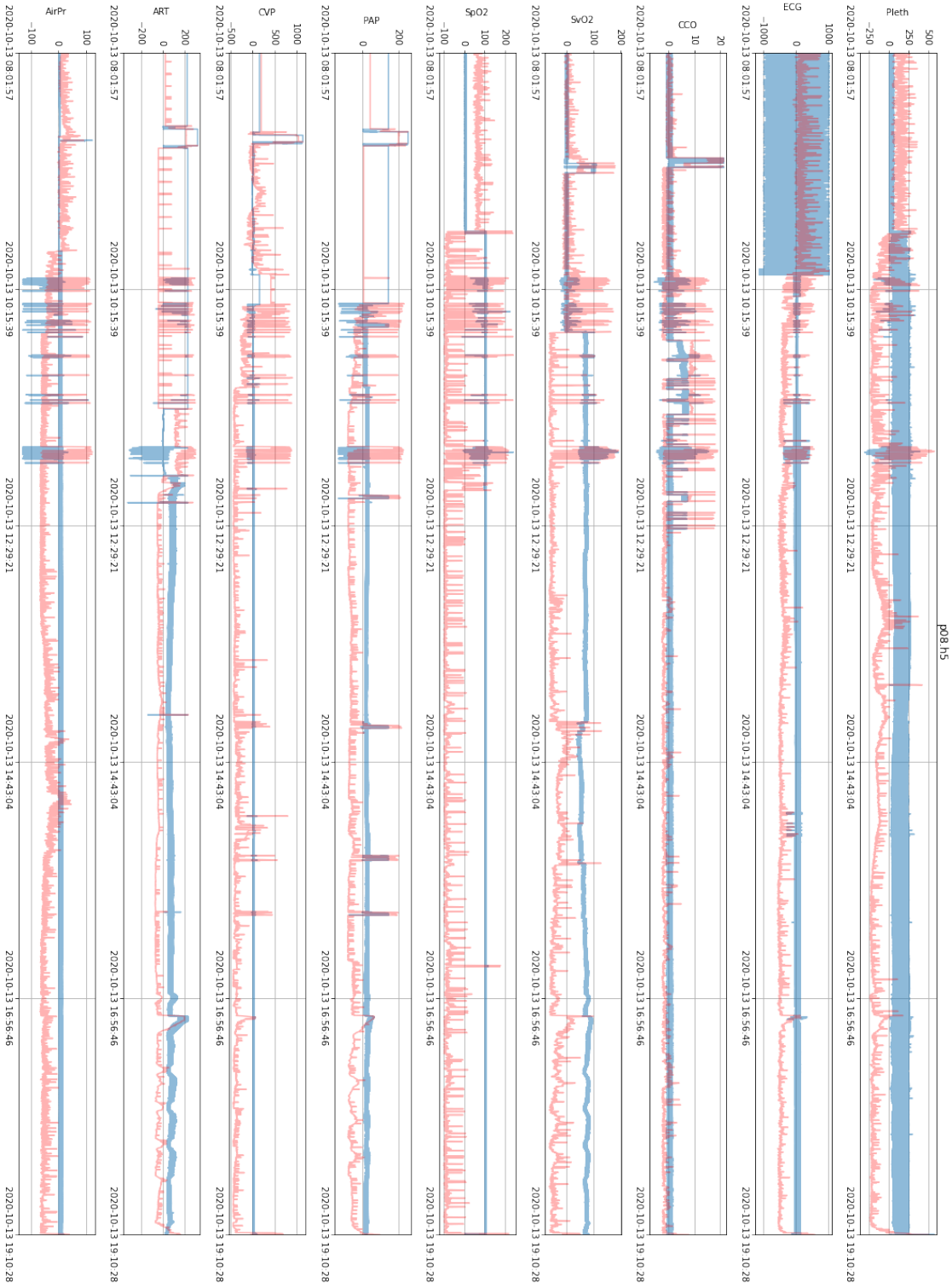


Figure 3.6: Results on all of the Pig08 waveforms (where IForest is fit individually on each signal). The Raw signal in shown in blue, and the scaled Anomaly score is shown in red.

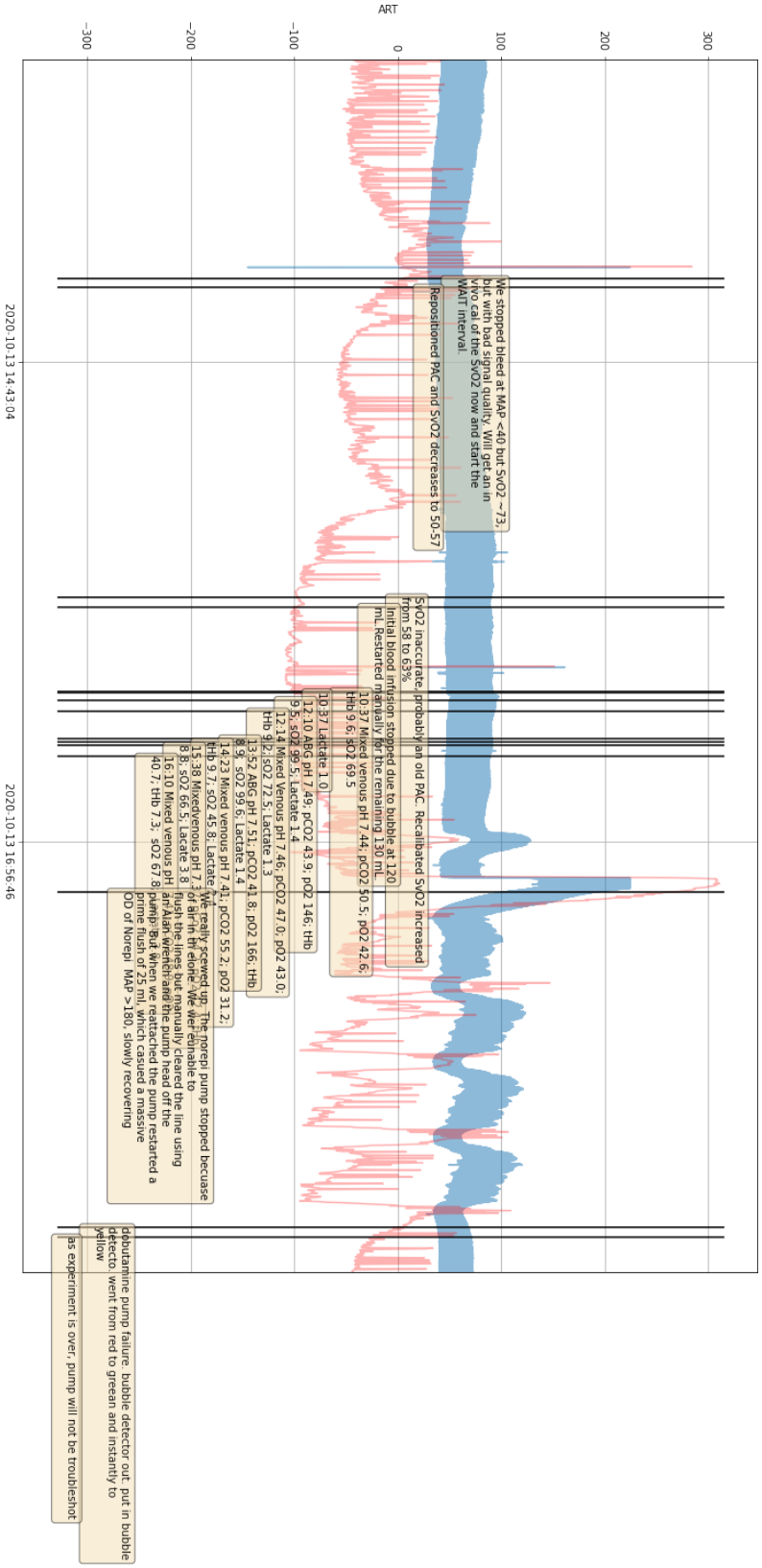


Figure 3.7: Results on ART (IForest is fit individually on each signal). The Raw signal in shown in blue, and the scaled Anomaly score is shown in red.

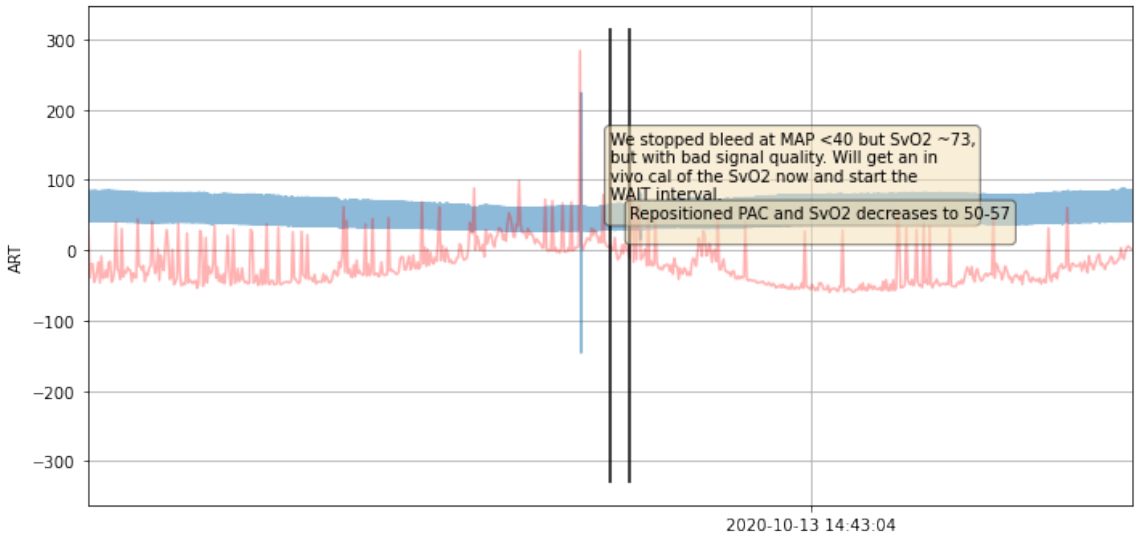


Figure 3.8: Zoomed in example annotation corresponding to a spike in the anomaly score.

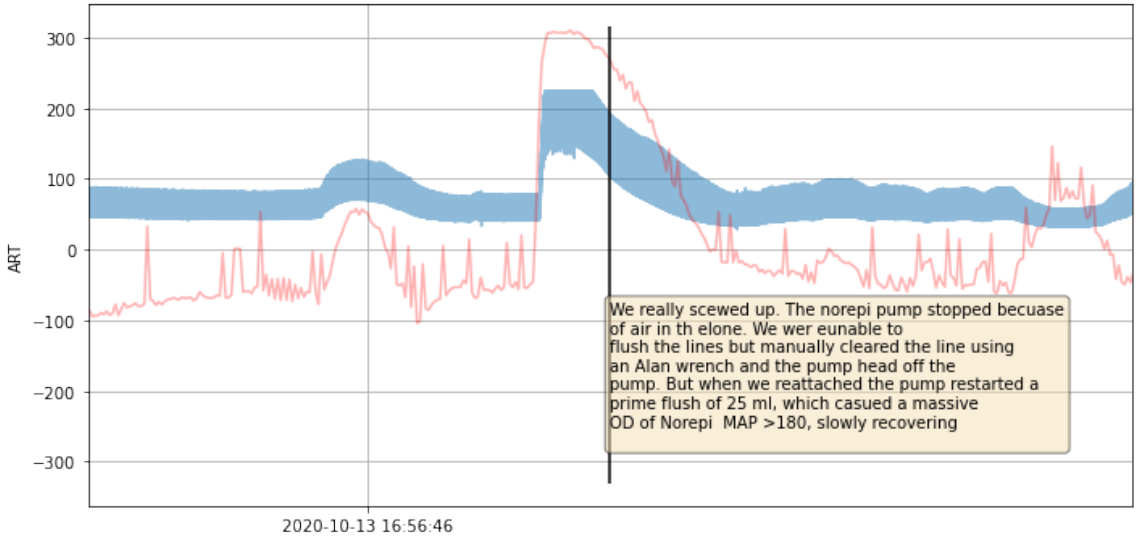


Figure 3.9: Zoomed in example annotation corresponding to a spike in the anomaly score.

From Figure 3.6, we see that obvious section of noise, such as the start (red box), has generally high anomaly score. However, we know from our clinician partners that ART is the most pertinent signal, so let us focus on that signal in particular.

From Figure 3.7, we see that many of the highly anomalous sections all correspond to real-world failures noted by the clinicians. For example, Figure 3.8 shows an example of a spike corresponding to the following annotation: *"We stopped bleed at MAP <40 but SvO2 73,but with bad signal quality. Will get an invivo cal of the SvO2 now and start the WAIT interval"*. Figure 3.9 shows an example of a spike corresponding to the following annotation: *"We really scewed up. The norepi pump stopped becuase of air in th elone. We were unable to flush the lines but manually cleared the line using an Alan wrench and the pump head off the pump. But when we reattached the pump restarted a prime flush of 25 ml, which caused a massive OD of Norepi MAP >180, slowly recovering"*.

This shows that the outlier detector on the SQI features is working properly for this case. Due to limitations, a deep analysis of each individual high anomalous area cannot be analyzed. However, this serves as a good first step for automatic detection.

3.2 ECG Denoising

We manually add noise from the MIT-BIH dataset to control the type and variety of noise added, without using nstdbgen, as it performs a custom procedure that is irrelevant to our use case.

3.2.1 Results on Open Source Dataset

We use 2 public datasets to test our methods. MIT-BIH arrhythmia and MIT-BIH arrhythmia noise stress test databases. The noise stress dataset is simply a collection of noise records that can be added onto clean signals from the MIT-BIH arrhythmia dataset. It contains noise records that exemplify baseline wander, or low frequency noise caused by motion. Muscle EMG artifact, or Electrical noise generated by muscle activity near the electrode. Finally, it also contains Electrode Motion artifact, or intermittent mechanical noise, which can mimic ECG beats.

For 2 clean records in the MIT-BIH dataset (118 and 119), we manually add 5 different types of noise (Electrode Motion artifact (EM), Baseline Wander (BW), Muscle Artifact (MA)), Gaussian Noise (GN), as well as a combination of all 4 previous noises at different levels of SNR (-6, 0, 6, 12, 18, 24). All the noise was

taken from the MIT-BIH noise stress test database except Gaussian Noise, which we added. Finally, we split ECG signals into windows of length 512, which is required by the autoencoder denoising method. After splitting data into training and testing sets, we end up with 9408 samples for each.

We will compare 3 main methods: Wavelet denoising, EMD denoising, and finally, Convolutional Autoencoder denoising as described in the Literature Review Section 2.2. As 2 additional baselines, we also compare using a simple 1D convolutional network and an LSTM to directly predict the clean signal from the noisy signal. The full implementation details of the adapted methods are in the file `denoising.py` in the released python package.

The evaluation Metrics consist of Mean Squared Error, Signal to noise ratio (SNR) measured in terms of decibels (dB), and finally, a Special Evaluation Metric specific for ECGs: ECGSNR [68]. For each method, the input will be a noisy signal, and the output is expected to be a cleaned version of that signal.

$$SNR = 10 \log_{10}(\frac{P_s}{P_n})$$

Where P_s is the power of signal in watts, and P_n is the power of noise in watts.

$$ECGSNR = 20 \log_{10}(\frac{S}{N})$$

Where S is average peak-to-peak amplitude of QRS complexes, and N is the average Root Mean Squared Error over 1 second windows of noise.

Noise Type	dB	Wavelet MSE	EMD MSE	LSTM MSE	CNN MSE	AE MSE
em	-6	0.0824	0.0684	0.0356	0.0237	**0.0187**
em	0	0.0527	0.0483	0.0236	0.0151	**0.0104**
em	6	0.0276	0.0337	0.0122	0.0094	**0.0067**
em	12	0.0101	0.0213	0.0059	0.0066	**0.0055**
em	18	**0.0028**	0.0171	0.0036	0.0048	0.0048
em	24	**0.0008**	0.0177	0.0030	0.0036	0.0046
ma	-6	0.0839	0.0853	0.0255	0.0222	**0.0109**
ma	0	0.0612	0.0681	0.0164	0.0148	**0.0076**
ma	6	0.0249	0.0370	0.0084	0.0096	**0.0063**
ma	12	0.0097	0.0246	**0.0046**	0.0058	0.0054
ma	18	**0.0026**	0.0187	0.0034	0.0045	0.0047
ma	24	**0.0010**	0.0189	0.0031	0.0042	0.0046
bw	-6	0.0794	0.0813	0.0176	0.0145	**0.0063**
bw	0	0.0608	0.0655	0.0114	0.0107	**0.0060**
bw	6	0.0378	0.0468	0.0093	0.0078	**0.0057**
bw	12	0.0178	0.0296	0.0061	0.0060	**0.0057**
bw	18	0.0077	0.0221	0.0037	0.0046	**0.0045**
bw	24	**0.0012**	0.0188	0.0029	0.0044	0.0051
gn	-6	0.0885	0.0941	0.0310	0.0314	**0.0272**
gn	0	0.0437	0.0549	0.0194	0.0168	**0.0144**
gn	6	0.0221	0.0364	0.0092	0.0080	**0.0066**
gn	12	0.0066	0.0212	**0.0046**	0.0055	0.0054
gn	18	**0.0026**	0.0184	0.0033	0.0044	0.0047
gn	24	**0.0011**	0.0174	0.0030	0.0038	0.0044
all	-6	0.0653	0.0685	0.0311	0.0359	**0.0229**
all	0	0.0487	0.0566	0.0192	0.0205	**0.0105**
all	6	0.0292	0.0412	0.0094	0.0110	**0.0065**
all	12	0.0113	0.0239	0.0048	0.0067	**0.0055**
all	18	**0.0035**	0.0187	0.0033	0.0046	0.0048
all	24	**0.0011**	0.0171	0.0029	0.0042	0.0045

Table 3.4: Mean Squared Error of the predicted output ECG vs the actual clean ECG for all 5 compared methods, at each type of noise and decibel.

Noise Type	dB	Wavelet SNR	EMD SNR	LSTM SNR	CNN SNR	AE SNR
em	-6	**3.3488**	1.6169	-0.1700	-0.4107	-0.3583
em	0	**2.5328**	0.8349	-0.2597	-0.5062	-0.2442
em	6	**2.0005**	0.2656	0.0340	-0.4531	-0.1242
em	12	**1.2296**	-0.4469	0.2928	-0.4731	-0.0892
em	18	**0.4950**	-1.2064	0.2416	-0.4928	-0.1079
em	24	0.1995	-1.7164	**0.2053**	-0.2091	0.0908
ma	-6	-7.5434	-8.9241	-2.6498	-1.1290	**-0.5028**
ma	0	-6.0144	-7.8036	-1.8899	-0.7218	**-0.2556**
ma	6	-1.9515	-3.7004	-0.9747	**-0.4304**	-0.0682
ma	12	**0.1385**	-1.6640	-0.0361	-0.3112	0.0185
ma	18	-0.3299	-2.1604	-0.0155	-0.5456	**-0.1067**
ma	24	-0.3845	-2.3505	**0.0163**	-0.4558	-0.0183
bw	-6	**3.9529**	3.0027	-1.5665	-0.7429	-0.3380
bw	0	**3.4577**	2.4977	-0.4652	-0.5566	-0.1517
bw	6	**2.7071**	1.4575	0.0051	-0.4428	-0.0817
bw	12	**1.8407**	0.7664	0.1834	-0.3855	0.0521
bw	18	**0.2811**	-1.5652	0.0798	-0.6456	-0.2407
bw	24	-0.1849	-2.1842	0.0833	-0.0887	**0.2695**
gn	-6	-8.4643	-9.7113	-1.8837	-1.4672	**-1.3796**
gn	0	-7.4742	-9.3032	-1.5990	-0.3594	**-0.3191**
gn	6	-4.6856	-6.6857	-0.9924	-0.4773	**-0.2842**
gn	12	-1.8117	-3.4528	-0.2708	-0.3207	**-0.0402**
gn	18	-0.8752	-2.7504	**-0.0539**	-0.5255	-0.1084
gn	24	-0.4862	-2.3357	**0.0305**	-0.6060	-0.1421
all	-6	-4.8649	-6.5392	-1.8861	-0.8658	**-0.8130**
all	0	-3.7288	-5.3820	-1.7129	-0.6127	**-0.3518**
all	6	-2.0208	-3.8272	-0.9819	-0.4693	**-0.1992**
all	12	-0.6469	-2.4161	-0.1546	-0.3931	**-0.0426**
all	18	-0.4290	-2.3001	**0.0157**	-0.5526	-0.0947
all	24	-0.2572	-2.2011	**0.0856**	-0.6421	-0.1514

Table 3.5: Signal-to-noise Ratio of the predicted output ECG vs the actual clean ECG for all 5 compared methods, at each type of noise and decibel. This default metric is not a good measure of model performance.

Noise Type	dB	Wavelet ECGSNR	EMD ECGSNR	LSTM ECGSNR	CNN ECGSNR	AE ECGSNR
em	-6	72.7573	72.6133	72.6275	72.5771	**72.8835**
em	0	72.7617	72.6890	72.6611	72.7470	**73.0096**
em	6	72.7309	72.6504	72.8036	72.8419	**73.0008**
em	12	72.8836	72.8321	72.9016	72.8804	**73.0042**
em	18	72.9656	72.9479	72.9744	72.9426	**73.0344**
em	24	73.0054	73.0144	73.0186	72.9853	**73.0464**
ma	-6	72.7239	72.5573	72.3124	72.5691	**72.9191**
ma	0	72.8369	72.6490	72.7029	72.8377	**73.0165**
ma	6	72.9650	72.8822	72.8908	72.8986	**73.0058**
ma	12	72.9889	72.9876	72.9559	72.9646	**73.0327**
ma	18	73.0051	73.0115	72.9972	72.9858	**73.0325**
ma	24	73.0295	73.0207	73.0220	72.9999	**73.0419**
bw	-6	73.0248	**73.0790**	72.9384	72.9564	72.8977
bw	0	73.0278	**73.0589**	72.9748	72.9814	72.9776
bw	6	73.0193	**73.0410**	72.9983	72.9488	73.0097
bw	12	73.0306	**73.0410**	73.0356	73.0073	73.0169
bw	18	73.0331	**73.0638**	73.0614	73.0014	73.0193
bw	24	73.0351	**73.0611**	73.0458	72.9586	73.0157
gn	-6	72.3538	72.4933	72.4382	72.2511	**73.0911**
gn	0	72.2418	72.3766	72.3094	72.3926	**73.1185**
gn	6	72.2117	72.2008	72.6663	72.7931	**73.0894**
gn	12	72.6332	72.6808	72.8384	72.8757	**73.0528**
gn	18	72.8911	72.8731	72.9220	72.8898	**73.0604**
gn	24	72.9486	72.9725	72.9605	72.9456	**73.0331**
all	-6	72.4245	72.1980	72.2271	72.2653	**72.8483**
all	0	72.3719	72.2060	72.3703	72.7184	**72.9845**
all	6	72.5968	72.5452	72.7478	72.8409	**73.0508**
all	12	72.8531	72.8247	72.8724	72.9013	**73.0652**
all	18	72.9441	72.9559	72.9469	72.9592	**73.0419**
all	24	72.9891	72.9968	73.0025	72.9604	**73.0114**

Table 3.6: ECG Signal-to-noise Ratio of predicted ECG vs the clean ECG for all 5 compared methods, at each type of noise and decibel. This specialized metric measures the noise ratio between the average peak to peak (QRS) amplitudes and the root mean square of the noise.

We see that, from both of the tables with reasonable metrics: Table 3.4 and Table 3.6, Convolutional Autoencoder generally performs the best at denoising ECG. In the Mean Squared Error case, it would be interesting to investigate methods where wavelet denoising is used for higher SNR and Convolutional Autoencoders are used for lower SNR. For the ECG SNR case, future work could focus on specifically preserving the amplitudes of the QRS signal, especially for the baseline wander noise type.

3.2.2 Case Study: Effects of ECG Denoising on Real vs Artifactual Alert Detection

One interesting application of ECG denoising is in the case of Real vs Artifactual Alert Detection, as in Section 3.1.2. Let the problem setup be the same as Section 3.1.2, except in the featurization. Since the 255 precomputed feature were computed with raw ECGs, and we unfortunately do not have access to the featurization functions, we instead do an ablation study with ONLY the SQIs. In this case, instead of passing the potentially noisy, raw ECG signal to the SQI featurizers, we pass a denoised version of the ECG instead.

This denoised version is obtained by resampling the raw ECG signal to a hz of 125 and splitting it into windows of length 512 for input to the CNN Autoencoder. As before, we use Random Forest with 10x patient level random splits for cross validations for all 3 types of alerts: SpO2, Heart Rate, and Respiratory Rate.

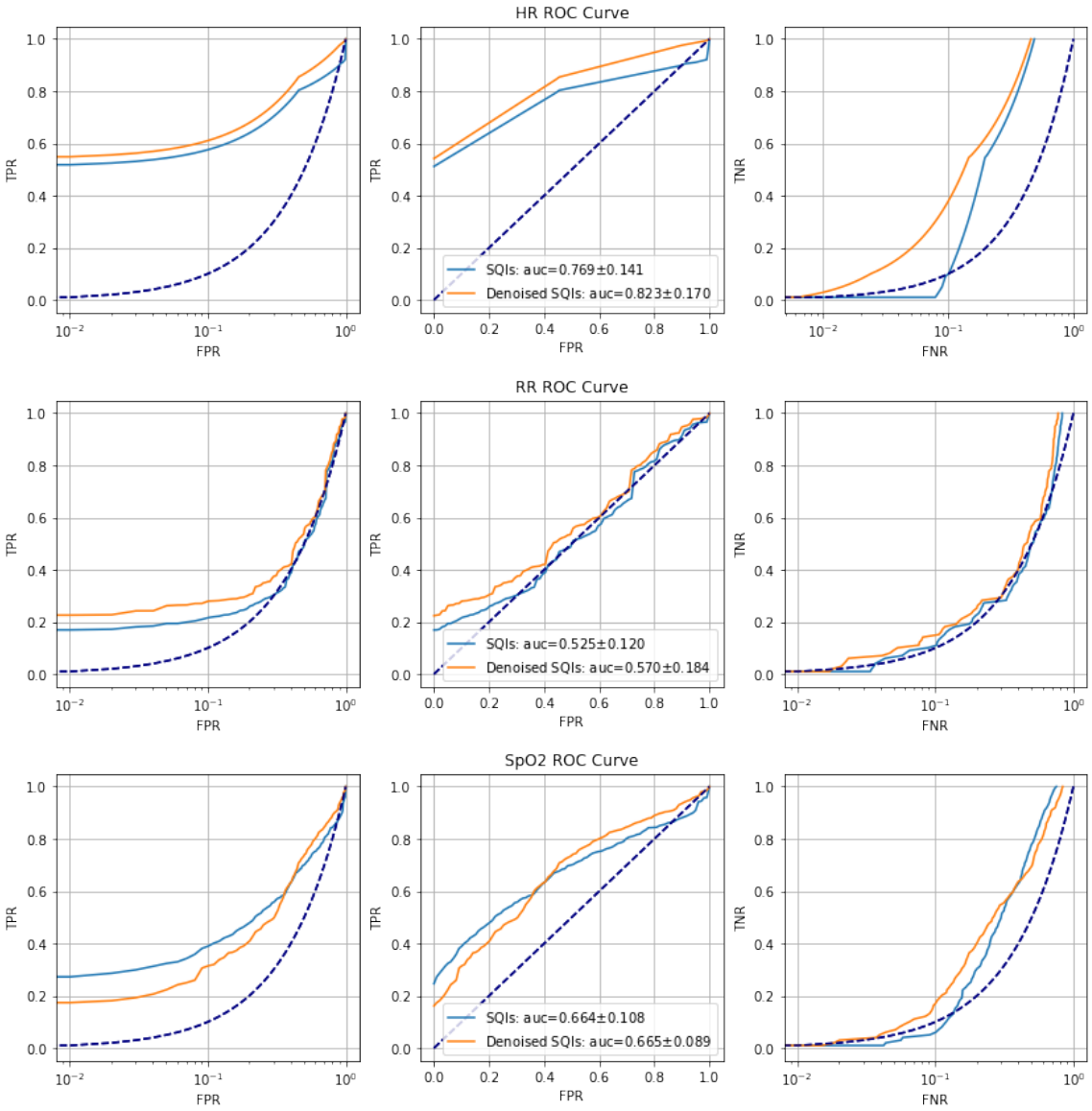


Figure 3.10: Plots of ROC Curves for all 3 alert types, with and without additional SQIs calculated on **denoised ECG**. For most alerts, we do improve AUC.

From Figure 3.10, we see that the AUC performance of the Random Forest is generally better compared to the results ran on Raw ECG signals in Figure 3.3, particularly for Heart Rate as well as Respiratory Rate. SpO2 AUC stays the same for both cases, which is consistent with the original results. This increase in performance indicates that our denoising algorithm is working, as it is removing noise from clean

signals, and failing to remove noise from noisy ones, increasing the difference between the 2 classes. This effect helps the Random Forest model more efficiently distinguish between real and artifactual alerts.

3.2.3 Results on Pig Data

Although there are only 25 pigs, each signal is of multiple hours long, and contains gigabytes of data. Additionally, we do not have a gold standard of a clean ECG. However, we are able to see visually, the difference between raw signals and cleaned outputs.

First, we resample the pig signals to 125 hz and windows of length 512 in accordance with our open source dataset. Then, we pass the original ECG signals through highpass Butterworth to remove slow drift and dc offset [61]. We also filter the signal with a powerline filter to remove it [61].

Finally, we pass the signal to our CNN Autoencoder for final processing. As a result, we obtain cleaned ECG signals that look like the following Figure 3.11.

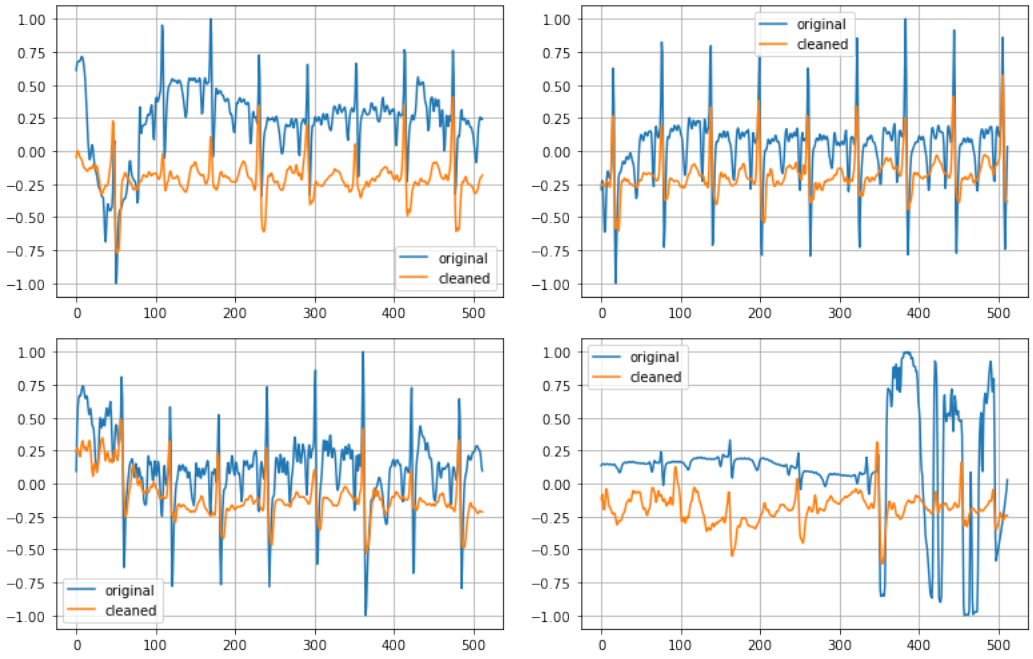


Figure 3.11: ECG Denoising Results on the Pig dataset. The blue is the original signal, and the orange is the cleaned signal.

From Figure 3.11, we can see that the cleaning process is able to at least remove some artifacts that occur in the original signal. For example, we see that in the first figure, the baseline wander component is removed. Additionally, we see that the CNN Autoencoder is able to preserve the peak to peak alignment quite well.

However, there are still issues to be solved, as in the bottom right figure, in cases of extreme noise, the algorithm may not be able to accurately denoise the signal. The CNN Autoencoder looses information in the first 2/3 of the time window. This could be related to hyperparameter choice (kernel size, filters), and the strict requirement of the neural network for a fixed length input time series. Section 3.3 addresses future work to tackle and potentially solve these issues.

3.3 Discussion

Signal Quality is vitally important in medical applications and machine learning for healthcare. However, these signals are often corrupted with different kinds of noises and artifacts. In this work, we studied common baseline methods in Signal Quality Indices, Outlier detection, and Signal Denoising. We demonstrated the capability of SQIs in reproducing 8 common ECG signal quality classification methods in the 2011 PhysioNet Challenge dataset. We further demonstrate the viability of SQIs by a case study in real-vs-artifact classification to reduce alarm fatigue. Additionally, after investigating 4 outlier detection methods, we validate our results on a closed-source bleeding dataset from the University of Pittsburgh Medical Center, and test our results on pig08. Furthermore, after testing 3 main ECG denoising approaches on the open source dataset MIT-BIH Arrhythmia, we also evaluated a case study showing that denoising improves real-vs-artifact detection. Finally, we conclude with examples of denoised ECG from the pig data. As mentioned previously, we will release an open-source implementation of all previously mentioned methods to serve as an accessible, open source, toolkit for signal quality analysis.

While this work does not directly focus on state of the art results, future work should be able to quickly reproduce our results, as a baseline to compare against. Ideally, this work serves as an accessible, open source, toolkit for signal quality analysis and ECG denoising. Additionally, future work may focus on different design choices for the CNN Autoencoder denoising model. Since kernel size, number of

layers, and dimensionality of deep learning models have a large impact in model capability, a more extensive evaluation of parameter choices is needed. Grid search, Bayesian Optimization, or other techniques could be used to search for the best parameter choices for the network. Furthermore, the current model is limited with a fixed size requirement of the input signal; future methods should seek to address this, perhaps with alignment via Dynamic Time Warping, or by recent Neural Network methods that do not require a fixed length input. Use of SQIs to measure signal quality in conjunction with the denoising model may also allow the user to quantify the expected improvement to the ECG signal.

Appendix A

ECG Review Continued

The following is a non-exhaustive list of the main approaches to filtering, detection, and managing ECG signal quality. Although no literature review can cover the entire breadth of relevant work, the most relevant papers and methods should be covered.

A.1 Literature Review

A.1.1 ECG Signal Quality

The following is a list of relevant and popular literature regarding ECG signal quality. Although there has been much work in this field, ECG signal quality metrics mainly fall into a few categories.

Rule-Based Methods

Since there are many obvious checks in ECG signal quality filtering, it is common for methods to operate with pre-programmed rules, such as flat-line detection, noise quantification, and more. The following are some examples of such approaches.

Powerline, Baseline Wander, Muscle Artifact Rules: Johannesen et al. [40] introduced an two-step algorithm to determine the quality of electrocardiograms (ECGs). First, they reject ECGs with macroscopic errors (signal absent, large voltage shifts or saturation). Second, three main sources of ECG noise were identified and

filtered. Powerline noise was filtered via a modified sinusoidal estimation-subtraction approach [51] and quantified via root-mean-square of the correction signal. Baseline wander was filtered by subtracting a cubic spline function fitted to the QRS 'onset' and quantified by root-mean-square of the cubic spline. Muscular noise was quantified by root-mean-square of the difference between the median beat and sinus beats. The 3 metrics were then summed to obtain a single quality metric. Evaluation was performed using the PhysioNet Challenge database (1500 ECGs) [94]. They achieved an overall classification accuracy of (92.3% for the training set and 90.0% for the test set.

Beat Detection Rules: Tat et al. [98] designed a pipeline of rules to determine the quality of ECG data including a Flat Line detection, the Tompkins et al. [75] real time QRS detection algorithm, thresholding of Heart Rate and standard deviation of amplitude. For the thresholds, common sense rules were applied, such as heart rate higher than 30 bpm and lower than 210 bpm. Additionally, they also accounted for misplaced electrodes, manually choosing the least noisy channel (V6) to obtain r-peaks. Evaluation was performed on the PICC dataset [94]. Finally, their method had a best score of 92% (accuracy).

Basic Signal Properties, Lead Crossing Points, and SNR Rules Hayn et al. [30, 31] implemented and evaluated four ECG rules. (A) First, basic signal properties such as signal amplitude, spike detection, and zero line detection were considered. (B) Second, the number of lead crossing points (a measure of the amplitude of the ECG signal) is calculated by plotting the leads. (C) Third, a quality measure was calculated from a combination of a) signal-to-noise ratio (amplitude of lowest QRS signal / highest amplitude of non-QRS signal), b) the maximum QRS amplitude, and c) regularity of the detected rhythm. (D) Fourth, the quality measure of the 2nd-worst channel was thresholded.

Finally, a rules based method combined each of the former conditions. Evaluation performed on the PICC dataset [94] showed that the best algorithm achieved an accuracy of 0.933 in the training set and 0.916 on the test set. While rules A and B may be accurate for real-time feedback during ECG self-recordings, QRS detection based measures (rule C) can further increase the performance if sufficient computing

power is available.

Filtering, Signal Noise, and Beat Residual Rules: Johannesen et al. [39] presented a rule-based algorithm based on flat line detection, QRS detection, and noise. First, lead-fail (flatline) in all leads was detected by checking the derivative of the signal. Second, high frequency noise (root mean square after QRS detection [76]) was thresholded as a rule. Third, signals with noise causing missed beats after QRS detection was discarded. Fourth, global low frequency noise (calculated after onset of QRS) is thresholded. Finally, a signal quality metric calculated on the residual of beats after subtracting the average is thresholded. Evaluation results on the PICC database [94] showed that the algorithm achieved a Se of 91% and a Sp of 85% for the training dataset, and an accuracy of 88% (Event 1) and 79% (Event 2) for the test dataset.

Filtering, Amplitude, Guassian Noise, and Rpeak Detetion Rules: Liu et al. [57] proposed 4 real-time signal quality assessment rules for ECG signal quality. First, a rule for flatline detection is calculated by checking the standard deviation of the signal. Second, a rule for detecting huge impulse is calculated by checking if the signal exceeded a certain threshold. Third, a rule for strong Gaussian noise is calculated via 2 methods. Sample entropy [85] of the signal was thresholded and power spectrums of the signal frequence over the noise frequency [9]: $\frac{\text{0.05 hz to 40 hz}}{\text{40 hz +}}$ was also thresholded. Fourth, a rule for R-peak detection error was calculated via a digital filter.

Based on the proceeding rules, signal quality indexes are calculated for both individual leads and on the combined set of leads, and are further thresholded. The results on the PICC dataset [94] yielded a sensitivity of 90.67% and a specificity of 89.78% on the test set respectively.

Rules + SVM Classification: Kužílek et al. [47] developed an algorithm based on simple rules and an SVM. Rules consisted of thresholds on variance, covariance, max, and range of the ECG signal. The SVM was trained on covariance and time-lagged covariance matrix elements. The output of both methods were then combined. Evaluation on the PICC dataset [94] yielded a training accuracy of .99 and a testing

accuracy of 0.836. Additionally, kurtosis, mean, or number of QRS signals detected were found to be not significant features in the SVM.

Simple ECG Rules: Moody et al. [65] proposed simple, real-time, heuristic rules including detecting flatlines, the overall ECG signal range, and the frequency of large changes (by measuring range in overlapping windows). Evaluation on the PICC dataset [94] with a combination of these rules yielded an accuracy of 91.3% for part A and /89.6% for part B.

ECG Rule List: Langley et al. [48] proposed 6 decision rules, applied in a cascaded manner. 1. flat baseline detection, 2. max value of signal, 3. high baseline drift (extracted by 6th order Butterworth lowpass filter), 4. low amplitude after baseline drift, 5. high amplitude after baseline drift, and 6. steep slope detection. Evaluation on the PICC dataset [94] with the combined algorithm yielded a score of 91.4% on the training set and 85.7% on the test set.

Flaline, Low Amplitude, Peak Artifact, Baseline Wander, High Frequency, Powerline Rules: Jekova et al. [38] presented a system of rules for ECG noise detection. It implements criteria for: 1. Flat line detection via the first derivative. 2. Detection of a low amplitude (LA) lead by scanning the peak-to-peak amplitude in the QRS boundaries after 4 Hz high-pass filtering. 3. Detection of peak artifacts (PA) by thresholding the output of 1 Hz high-pass filter for specific slopes and amplitudes. 4. Detection of baseline wander by thresholding the mean of 1 Hz low-pass filter. 5. Detection of electromyographic and other high-frequency (HF) noises by computing the SNR of the output of 20 Hz high-pass filter. 6. Detection of powerline interference by scanning the outputs of two band-pass filters BP50 (48–52 Hz) and BP60 (58–62 Hz).

In total, this yields 13 adjustable thresholds for amplitude and slope criteria which are evaluated in adjustable time intervals, as well as number of leads. Evaluation on the PICC dataset showed a sensitivity of 98.7% an a specificity of 80.9%.

Signal Quality Index (SQI / SQA) Methods

Flat Line, Missing Channel, Amplitude, and Correlation SQIs: The winner of the Physionet Challenge 2011 (PICC dataset) [94], Xia et al. [104] proposed a 5-stage algorithm for ECG signal quality detection. First, if the ECG has any flat channels, then it is noisy. Second, if the ECG has any missing channels, then it is noisy. Third, if the signal has an amplitude that is too high or small, it is also noisy. Fourth, cross correlation is computed for different channels of the ECG, and if it is smaller than a threshold, it is determined to be noisy. Fifth, a similar calculation is performed for self-correlation. The authors also proposed a secondary method, where they constructed a 12x12 matrix where diagonal elements are the quality of the 12 channels of ECG, and the off-diagonal elements are the quality of correlations between channels. The authors state that in the best case scenario, this matrix is all zeros (since correlation ranges from [0,1], they assume that the authors are using 1-correlation). The authors the use the largest absolute eigenvalue of this matrix as a measure of quality of the ECG.

Finally, the authors were able to achieve a training accuracy of 86% on the using the first method and a training accuracy of 93.5% and testing accuracy of 90.0% using the second method.

The dataset used are the Physionet/Cinc Challenge 2017 [14] as well as the 2011 challenge dataset [26, 94]. Accuracies of with an accuracy of 92.67% on the 2017 challenge and 94.67% on the 2011 challenge datasets are achieved by the described method. Although the SVM method proposed by Clifford et al. [15] is still the best, the authors emphasize the interpretability of this method while still obtaining good performance.

Normalized Area Difference Between Successive QRS SQI Wang et al. [102] proposed a signal quality measure based on the normalized area difference between successive QRS complexes.

$$\text{mismatch}(x, y) = \frac{\sum_i |x(i) - y(i)|}{\sum_i |x(i)| + \sum_i |y(i)|}$$

Where the mismatch is 0 if the 2 QRS waveforms are identical or 1 if there is no overlap (e.g. $x = -y$). After a few processing steps such as aligning the beats and computing histograms, the signal quality is determined by the rise of the leftmost histogram bin curve. Evaluation was performed on the MIT-BIH two-channel arrhythmia database [68]. Results on 44 records showed that the method identified 22 records where one of the two leads had much better signal quality than the other lead. Furthermore, the method also improves arrhythmia detection, as the average PVC false positive rate was 0.47% for high quality leads compared to the average PVC false positive rate of 2.56% for low quality leads.

Modulation Spectral SQI: Falk et al. [19] proposed a ECG quality index MS-QI, based on the modulation spectral signal representation. This representation quantifies the rate of change of ECG spectral components, which are shown to be different from the rate of change of typical ECG noise sources. MS-QI is akin to a SNR measure. The authors identified bins in the frequency components that corresponded to ECG modulation energy and compared it to the remaining modulation energy as a ratio.

MS-QI was tested on 4 datasets: 1) synthetic ECG signals corrupted by varying levels of noise from the MIT-BIH Noise Stress Test [68], 2) a real-world dataset of single-lead ECG recorded using the Hexoskin garment over sitting, walking, and running, 3) the PICC dataset [94], and 4) MIT-BIH Arrhythmia Database [67]. Experimental results showed that MS-QI performed more reliably than Kurtosis and in-band (i.e., 5–40 Hz) to out-of-band spectral power ratio [12].

114 SQIs: Li et al. [54] presented a novel framework for false alarm (FA) reduction using a machine learning approach to combine up to 114 signal quality and physiological features extracted from the electrocardiogram, photoplethysmograph, and more. The features extracted from ECG include heart rate, two metrics of inter-channel and inter-algorithm comparisons of two QRS detectors, kurtosis, spectral distribution of ECG. Additionally, a fusion of the ECG metrics were calculated [52]. A genetic algorithm [49] was used for feature selection, and a Relevance Vector Machine [100] (a sparse Bayesian model) was trained and evaluated on the MIMIC-II [88] dataset. Results showed that the implemented algorithm would not suppress any true alarms with high confidence, while maintaining a moderate suppression rate for

false ventricular tachycardia alarms.

Multichannel Adaptive Filter SQI: Silva et al. [95] presented a new generic point-by-point signal Multichannel Adaptive Filter (MCAF) SQI based on adaptive multichannel prediction [93] that does not rely on ad hoc morphological feature extraction. The MCAF prediction algorithm consisted of a bank of M gradient adaptive Laguerre lattice (GALL) filters [20] followed by a Kalman filter [29]. Each GALL attempts to reconstruct the target signal from the input channels. The SQI is calculated by comparing the predicted signal to the actual signal, along with a flatline detection mask. Evaluation on the MIMIC-II dataset [88] yielded a AUROC of 0.86 (PPG), 0.82 (ABP), and 0.68 (ECG). Additionally, the SQI is monotonically related to signal-to-noise ratio (simulated by adding white Gaussian noise).

SQIs Applied to ECG SmartVest: Liu et al. [58] developed a novel wearable 12-lead ECG device that contains a real-time signal quality assessment (SQA) and lightweight QRS detection system.

Other Machine Learning Methods

SNR of the QRS template: Quesnel et al. developed a method for calculating metrics based off of the SNR of the QRS template [80, 81]. The method [80] first calculates an average PQRST complex, then compares each PQRST complex to it. The signal to noise ratio (SNR) for each beat is calculated as

$$SNR_{window} = \min_{i=1\cdots N} SNR_i = \min_{i=1\cdots N} 20\log\left(\frac{RMS(PQRST_{average})}{RMS(PQRST_i - PQRST_{average})}\right)$$

for beat i where RMS is the root mean square. The minimum SNR_i used for the analysis window. Evaluation on noise from the PhysioNet MIT-BIH Noise Stress Test Database [68] and a clean ECG signal [37] yielded a correlation coefficient of 0.89 (estimate vs real SNR), suggesting that the proposed SNR estimation algorithm produces results that reflect the actual SNR for a given ECG signal.

Anomalous Amplitude and Frequency Filtering: Redmond et al. [83] described an algorithm for automatically marking ECG recordings for obvious artifact

detection. The algorithm consists of 3 signal masks. The first is a Rail Contact Mask that segments a 1 second window around anomalous amplitudes. The second is a High Frequency Mask that notch filters out 50hz and high-pass filtered at 40 Hz. The third is a Low Power Mask that performs band-pass filtering with a passband of 0.7 - 33 Hz. Evaluation on 4751 single lead-I ECG recordings from 24 home-dwelling patients showed a sensitivity of 89% and a specificity of 98%.

Clustering and Kernel Density Estimation Redmond et al. [84] introduced an algorithm to determine the quality of single-lead electrocardiogram (ECG) recordings obtained from telehealth patients. After removing gross movement artifacts by masking and filtering [83], detected beats (using Pan Tompkins [75]) were first clustered using K-means. Features are then calculated on the clusters, such as mean ratio of vector error norm to mean cluster vector norm, and so on. Also, correlations such as mean correlation of entire beat with clustered beat shape are also calculated. Parzen window classifier (Kernel Density Estimation) was then applied to estimate the remaining ECG signal quality. Evaluation was performed on 300 short ECG recordings were manually annotated to identify movement artifact, QRS locations and signal quality (discrete quality levels). The classifier yielded an accuracy of 78.7% , similar to the gold standard annotation (accuracy = 70–89.3%).

Energy-Concavity Index (ECI) and a correlation-based metric: Naseri et al. [72] proposed a three stage algorithm including preprocessing, energy-concavity index (ECI) analysis, and a correlation-based final step. First, the preprocessing removes high frequency disturbances by gaussian smoothing [32, 71]. It then removes baseline wander by adaptive smoothing filtering. Then, an energy-based signal quality metric (energy-concavity index) was calculated. The signal concavity was calculated as mean slope of lines drawn from the centre of the window to all the points of the window samples. Then, the signal concavity was normalized. Finally, energy of the normalised concavity curve is calculated by gaussian smoothing filtering. A correlation-based quality measurement is also calculated using the correlation between ECG leads estimated by applying a suitably trained neural network.

Evaluation was performed on the PICC dataset [94] as well as a personal dataset [24] of 25 3-lead, 1-hr long, ECGs. On the personal dataset, the combined ECI and

correlation achieved a sensitivity (Se) of 77.04% with a positive predictive value (PPV) of 90.53% for detecting high-energy noise. On the PICC dataset, the combined metrics achieved a final score (accuracy) of 93.60%

Empirical Mode Decomposition Methods: Lee et al. [50] presented a two-stage method for the detection of motion and noise (MN) artifacts. The first stage applies empirical mode decomposition (EMD) [35] to isolate HF components of the signal. Specifically, the first-order intrinsic mode function (F-IMF) from the EMD is used. This goal is to isolate the artifacts' dynamics as they are largely concentrated in the higher frequencies. The second stage calculates Shannon entropy, mean, and variance on the F-IMF time series. These 3 features are then thresholded for classification of noisy / clean. Training was performed on ECG (Holter Monitor) data from 15 healthy subjects to derive threshold values associated with these statistical measures. Evaluation on 30 additional subjects show a sensitivity of 96.63% and specificity of 94.73%.

Wavelet Transform and Hidden Markov Model: Sangaiah et al. [90] proposed a three phrase framework to analyze ECG. First, the ECG signal quality is filtered to reduce noise. Motion artifacts and baseline wander is removed by an IIR elliptic high pass filter of order 3 and with a cut-off frequency of 0.5 hz. Power line interference is removed by a 60 hz IIR notch filter. Second, a wavelet transform [2] is used to extract features. Third, a hidden Markov model (HMM) is used for cardiac arrhythmia classification.

The features extracted include the maximum, minimum, mean, standard deviation, and median of each wavelet coefficient. Evaluation on 45 ECG records in MIT BIH arrhythmia database [67] and MIT BIH noise stress test database [68] yields an accuracy of 99.7 % with a sensitivity of 99.7 % and a positive predictive value of 100 % for cardiac arrhythmia classification.

Autocorrelation Features and RUSBoost (Ensemble Boosting Classifier for Imbalanced Data) Moeyersons et al. [64] presented an that uses features from the autocorrelation function to classify signal quality using RUSBoost, a probabalistic, tree-based, boosting, ensemble method that works well for imbalanced datasets. First,

the ECG was filtered by a zero phase, 2nd order high-pass and 4th order low-pass Butterworth filters with cut-off frequencies at 1 Hz and 40 Hz, respectively to remove baseline wander and high frequency noise. Second, ECG windows were passed thru the autocorrelation function, and first min, max amplitude, and similarity was extracted for classification via RUSBoost. The probability of the clean class is used as a SQI. Evaluation was performed on 3 datasets: a sleep dataset [101], the PhysioNet 2017 Challenge dataset [14], and a stress dataset [36]. AUC values between 0.988 and 1.000 were obtained when training the model on the Sleep dataset and evaluating on the other datasets.

A.2 Open Source Datasets

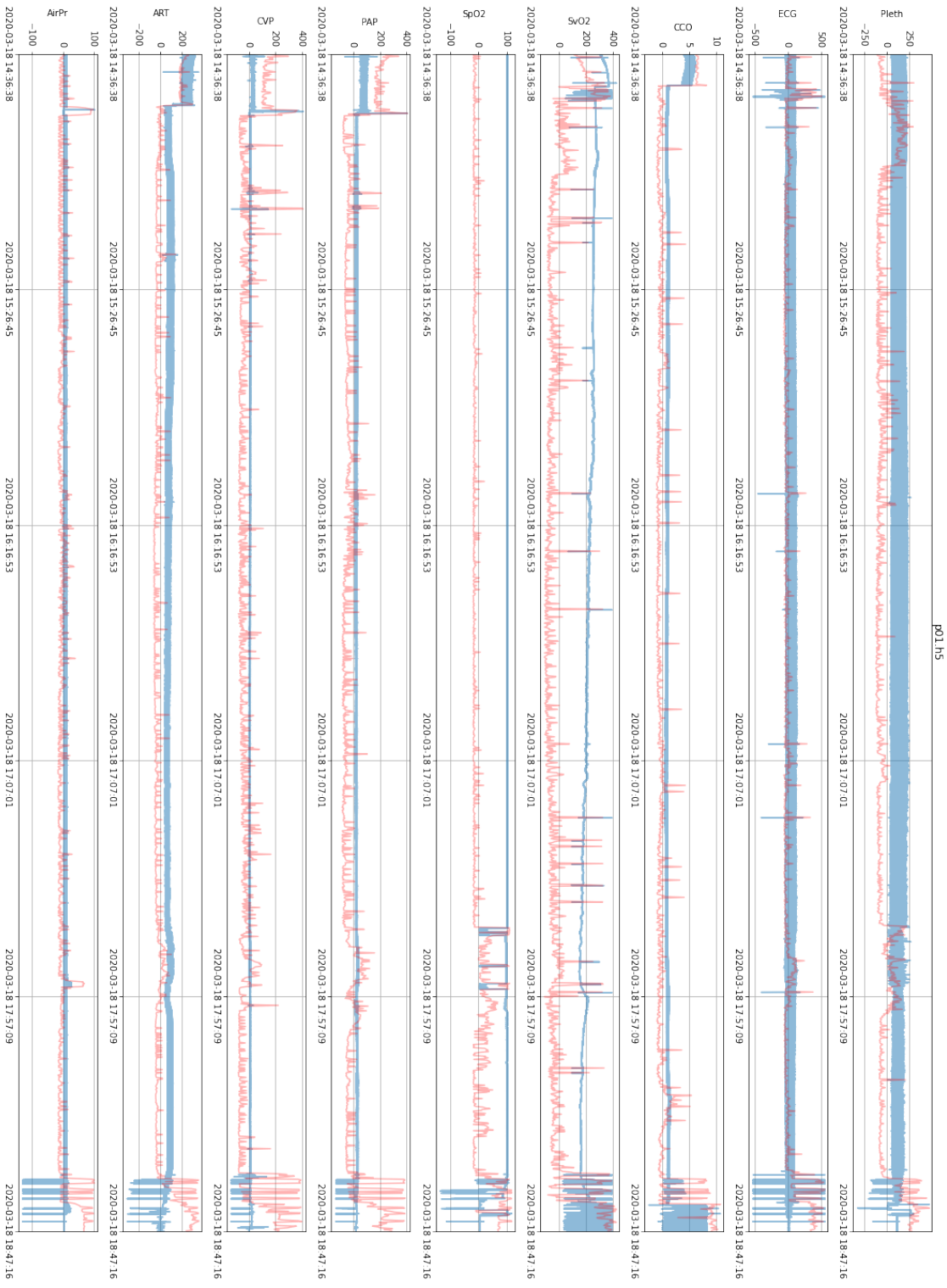
The MIMIC-II database (Multiparameter Intelligent Monitoring in Intensive Care II) [88] MIMIC is a very large database of patients from the ICU. The updated versions, MIMIC-III [42] and MIMIC-IV [41] are the more commonly accepted and used version, and both are accessible via PhysioNet [26] after a data usage agreement. The database consists of 25,328 intensive care unit stays. Detailed information about intensive care unit patient stays, including laboratory data, therapeutic intervention profiles such as vasoactive medication drip rates and ventilator settings, nursing progress notes, discharge summaries, radiology reports, provider order entry data, International Classification of Diseases, 9th Revision (ICD-9) codes, were collected. For certain patients, high-resolution vital sign waveforms and chest x-rays were also collected. Data were automatically deidentified to comply with Health Insurance Portability and Accountability Act (HIPAA) standards and integrated with relational database software to create electronic intensive care unit records for each patient stay.

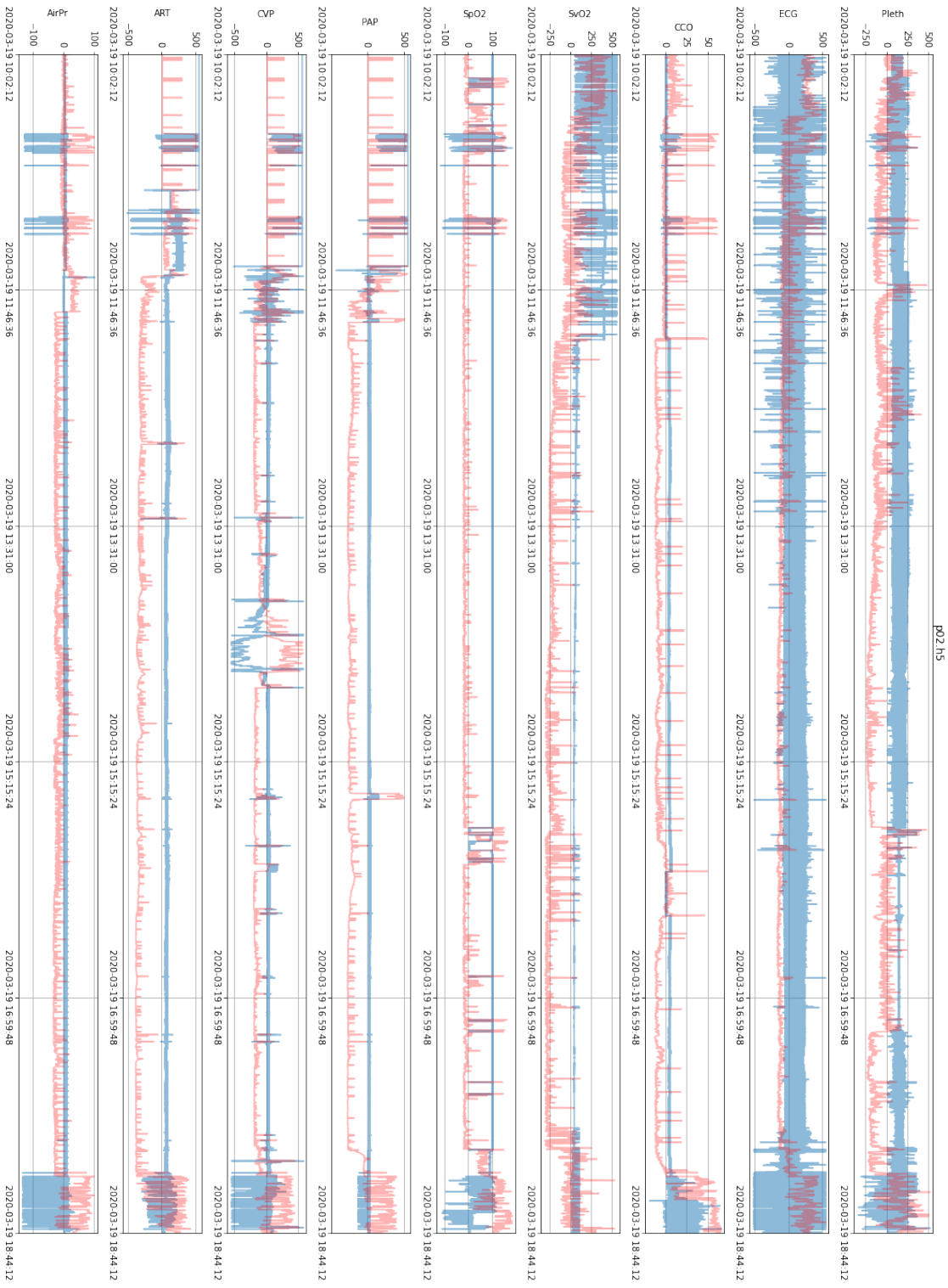
ECGSYN This realistic ECG waveform generator [63] is able to synthesize an ECG signal with user-settable mean heart rate, number of beats, sampling frequency, waveform morphology (P, Q, R, S, and T timing, amplitude,and duration), standard deviation of the RR interval, and LF/HF ratio (a measure of the relative contributions of the low and high frequency components of the RR time series to total heart rate variability). Using a model based on three coupled ordinary differential equations, ECGSYN reproduces many of the features of the human ECG, including beat-to-beat

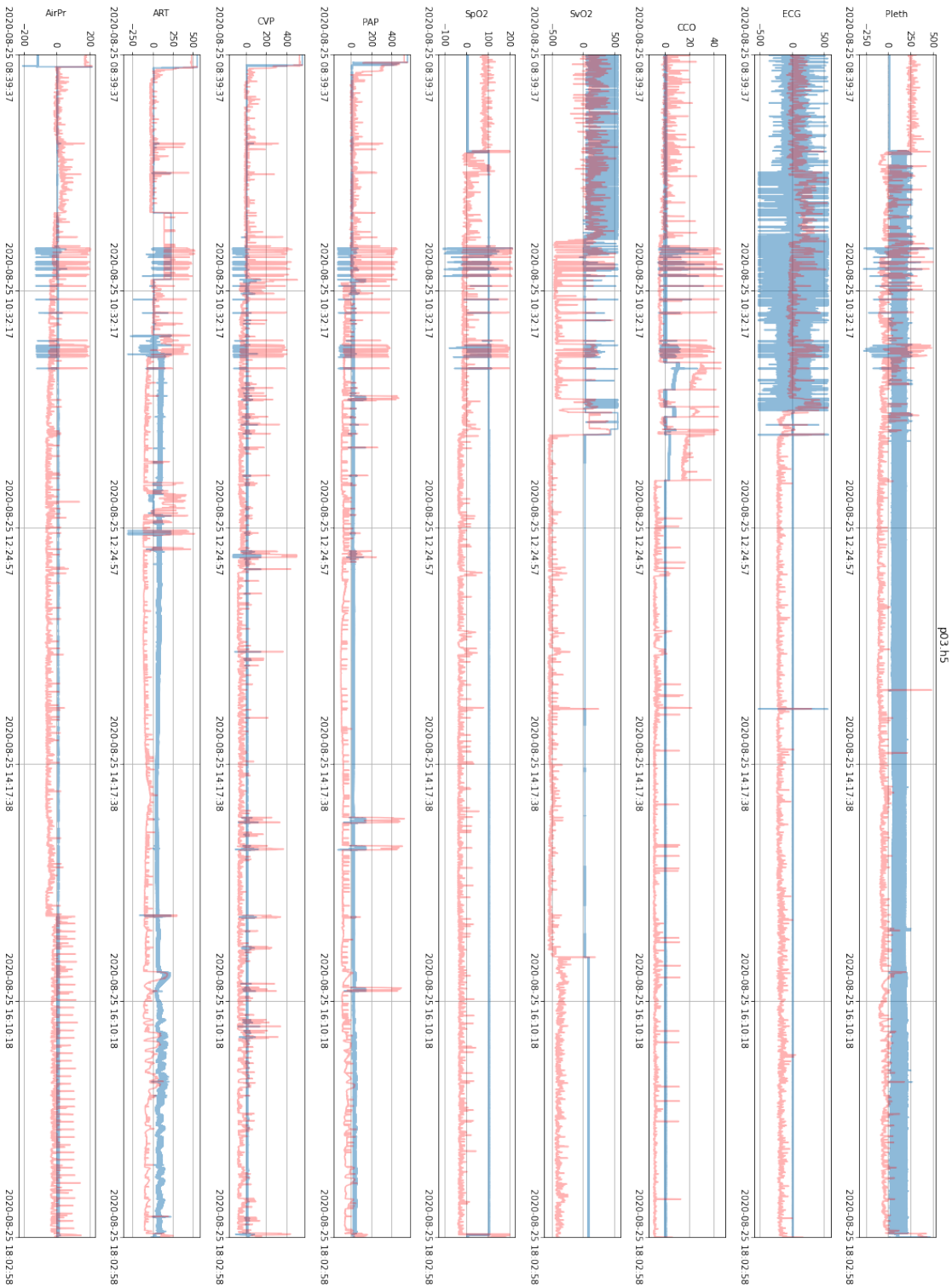
variation in morphology and timing, respiratory sinus arrhythmia, QT dependence on heart rate, and R-peak amplitude modulation. The output of ECGSYN may be employed to assess biomedical signal processing techniques which are used to compute clinical statistics from the ECG. An open source implementation is available on Physionet [26] as well as in the Neurokit Python Package [61].

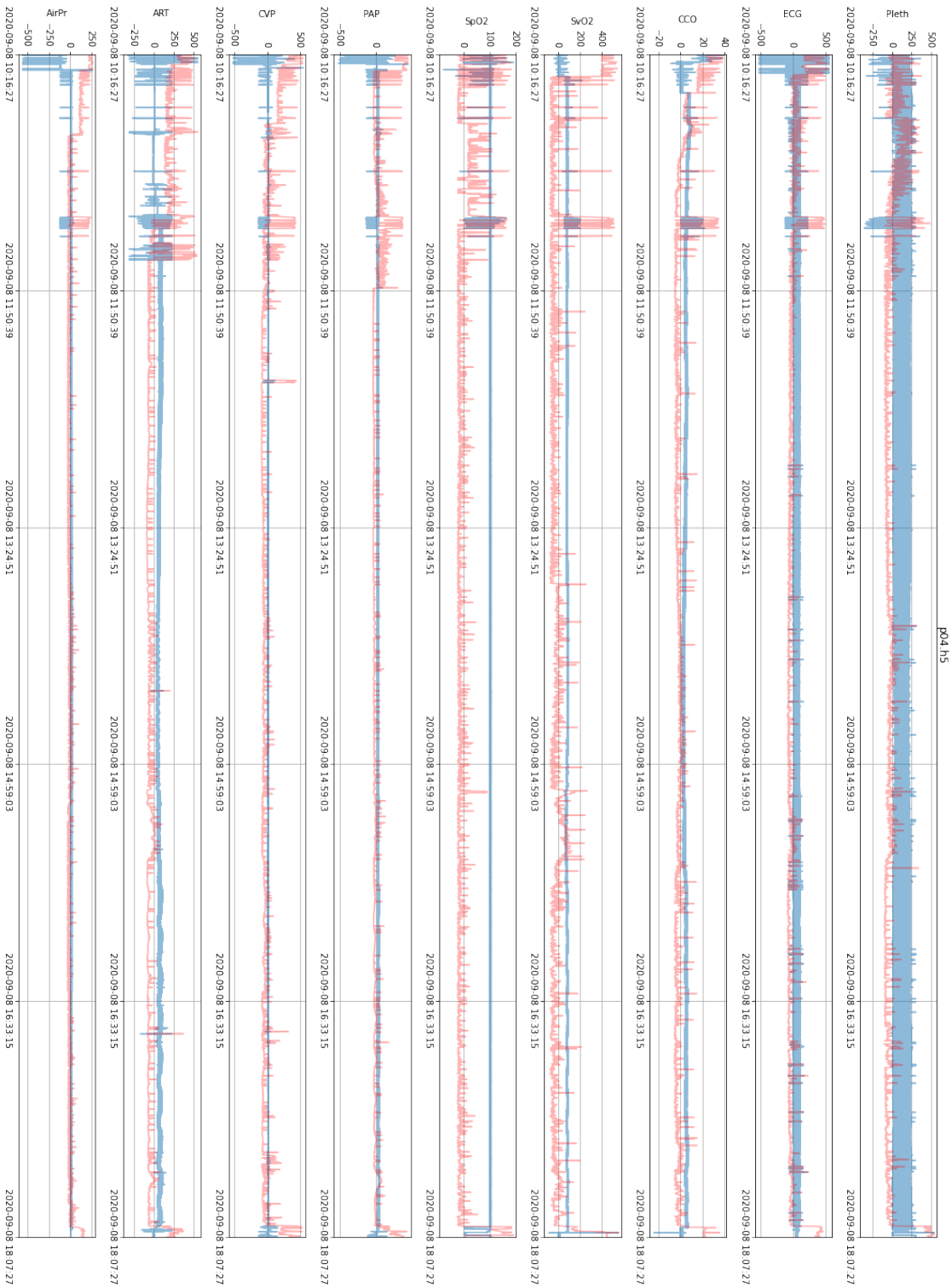
Appendix B

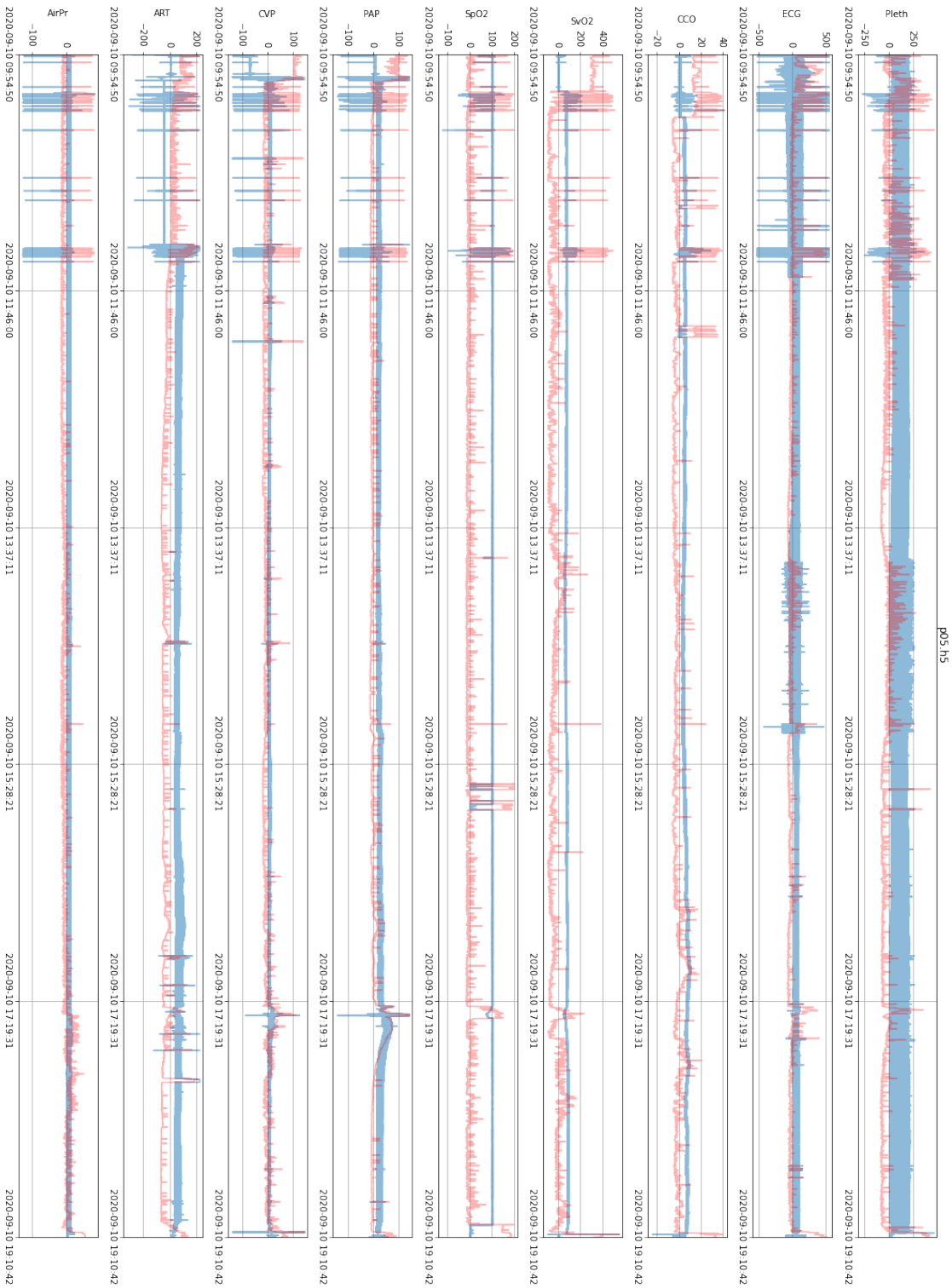
Additional Pig Signal Quality Outlier Detection Results

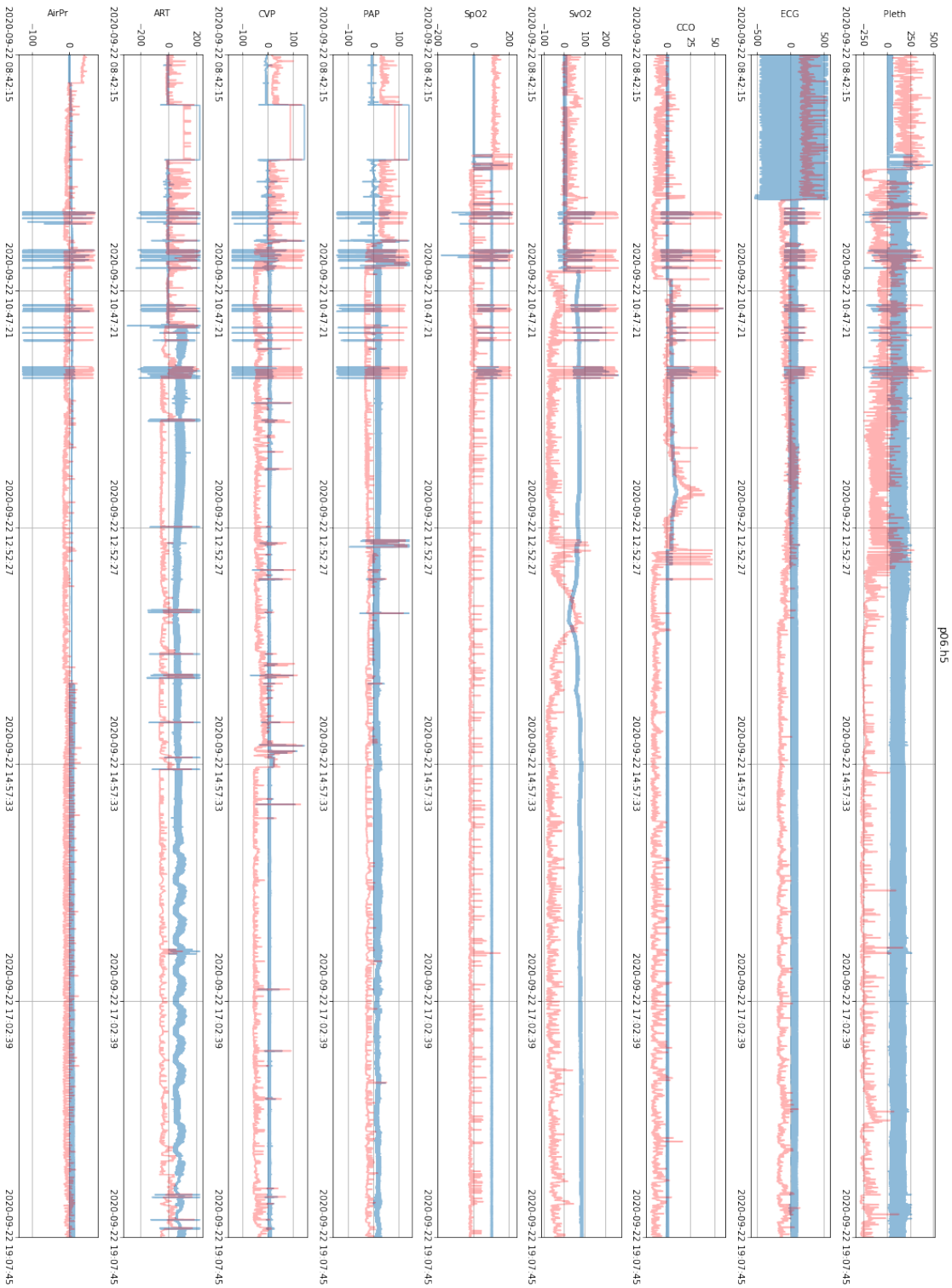

p01.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-03-18 14:36:38
2020-03-18 15:26:45
2020-03-18 16:16:53
2020-03-18 17:07:01
2020-03-18 17:57:09
2020-03-18 18:47:16

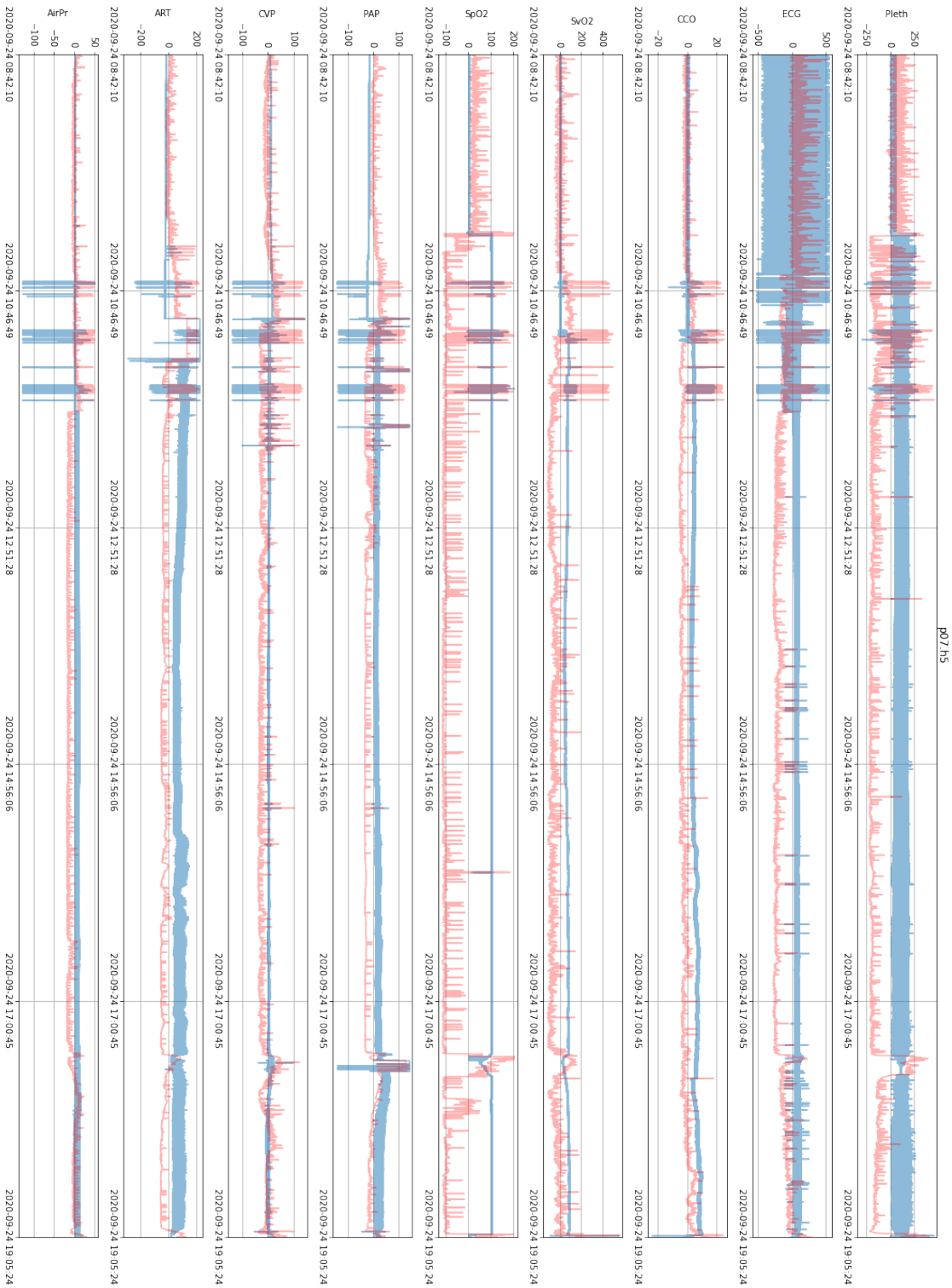

p02.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-03-19 10:02:12
2020-03-19 11:46:36
2020-03-19 13:31:00
2020-03-19 15:15:24
2020-03-19 16:59:48
2020-03-19 18:44:12

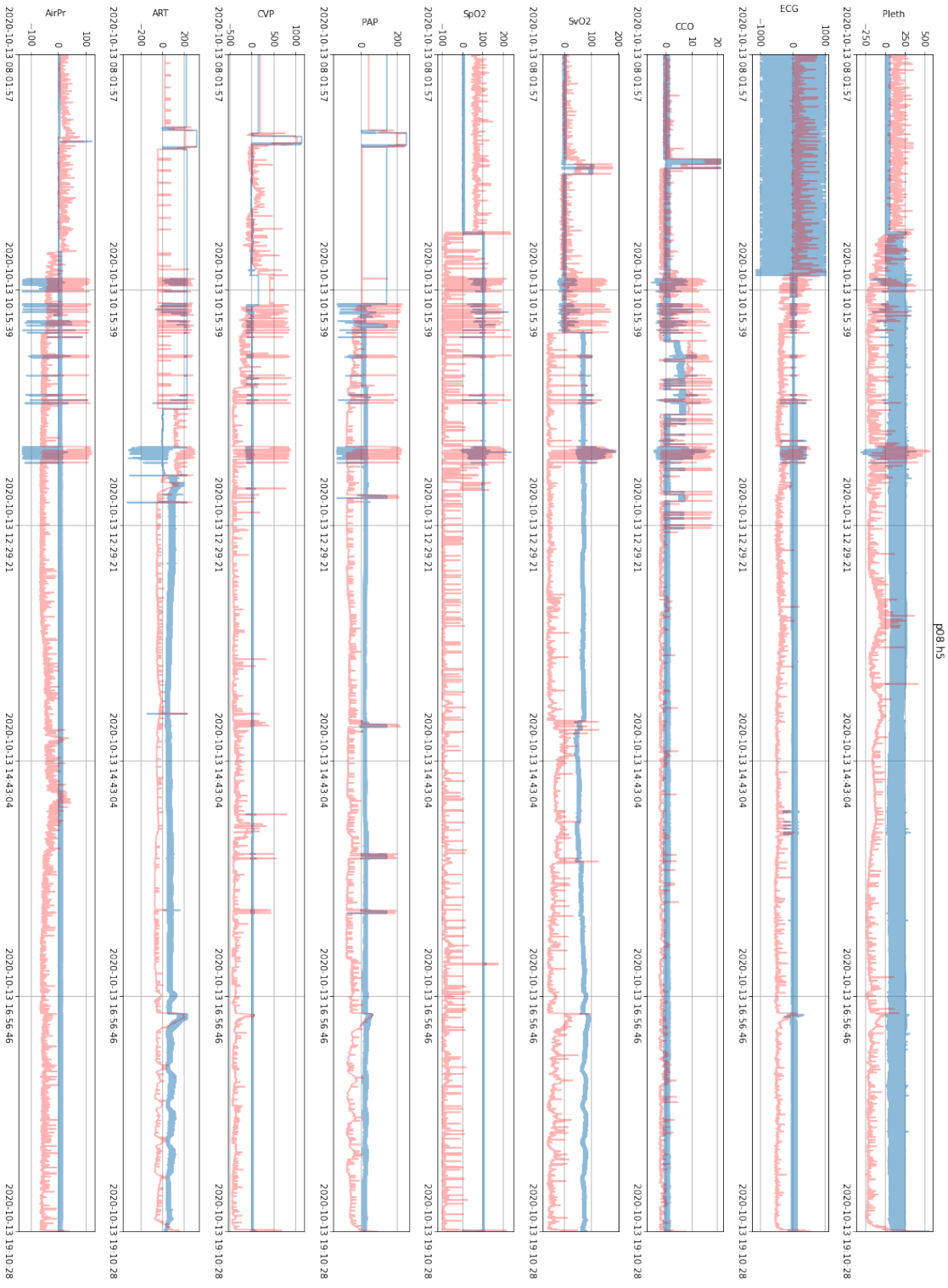

p03.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-08-25 08:39:37
2020-08-25 10:32:17
2020-08-25 12:24:57
2020-08-25 14:17:38
2020-08-25 16:10:18
2020-08-25 18:02:58

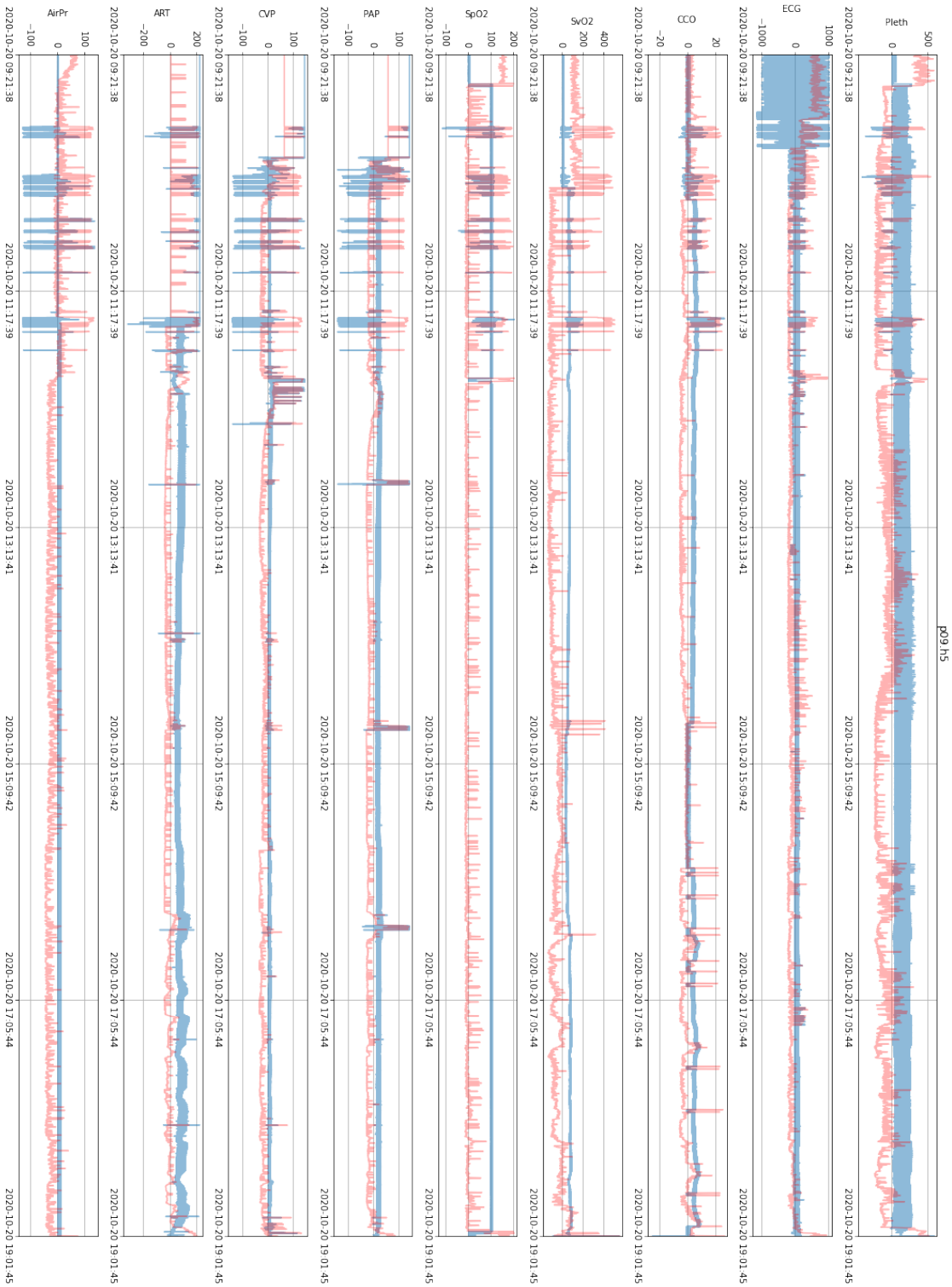

p04.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-09-08 10:16:27
2020-09-08 11:50:39
2020-09-08 13:24:51
2020-09-08 14:59:03
2020-09-08 16:33:15
2020-09-08 18:07:27

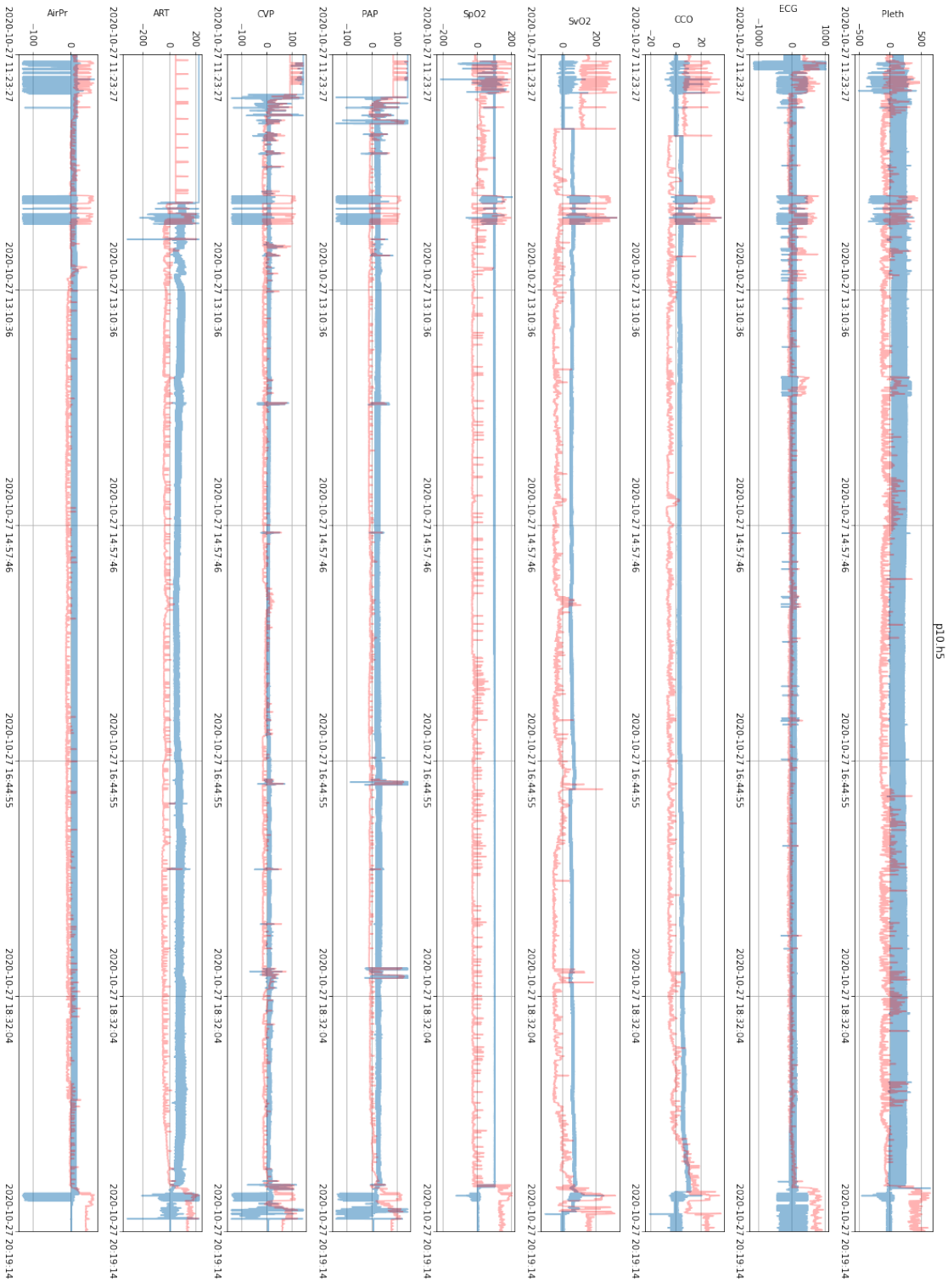

p05 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-09-10 09:54:50
2020-09-10 11:46:00
2020-09-10 13:37:11
2020-09-10 15:28:21
2020-09-10 17:19:31
2020-09-10 19:10:42

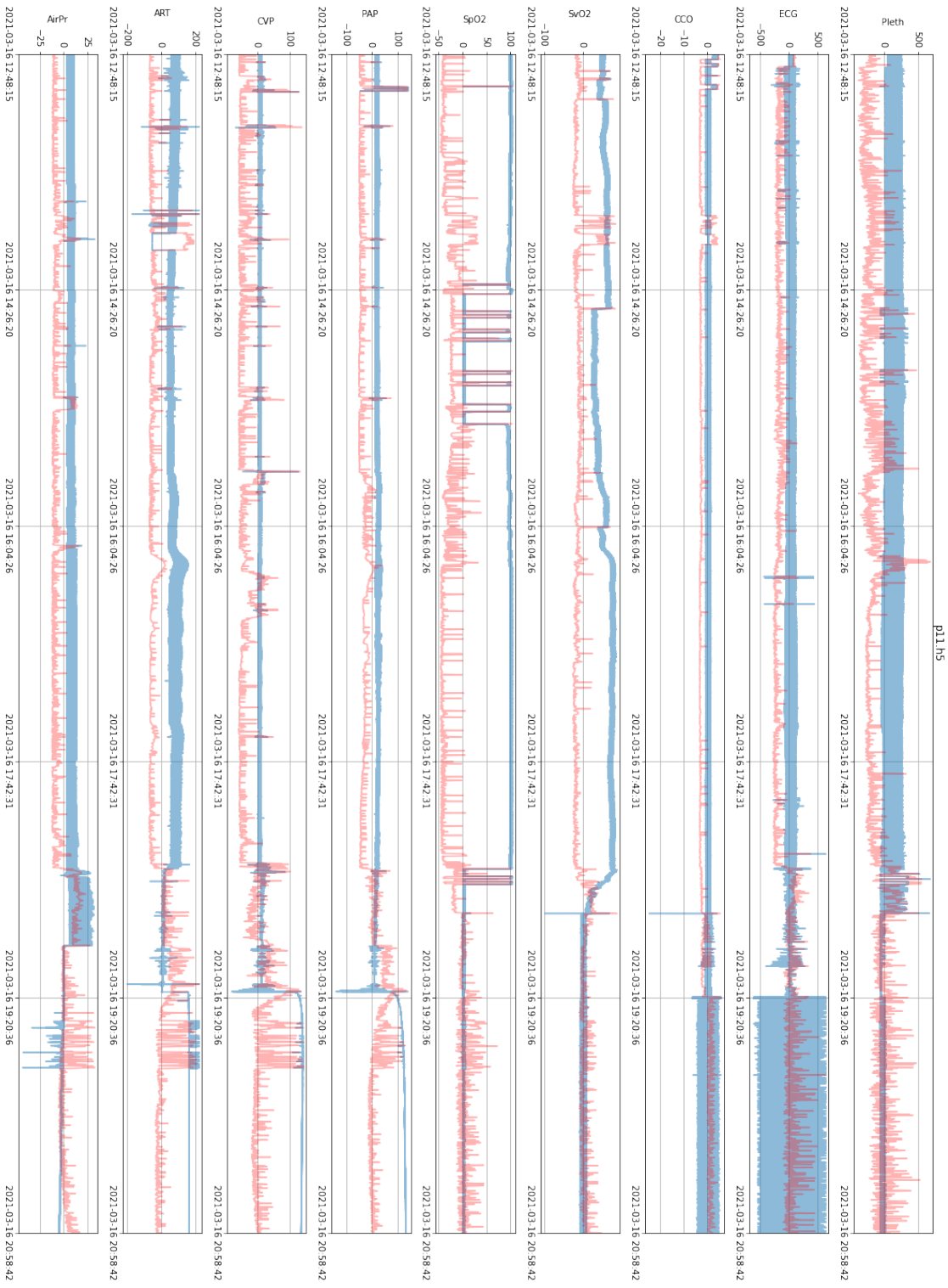

p06.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-09-22 08:42:15
2020-09-22 10:47:21
2020-09-22 12:52:27
2020-09-22 14:57:33
2020-09-22 17:02:39
2020-09-22 19:07:45

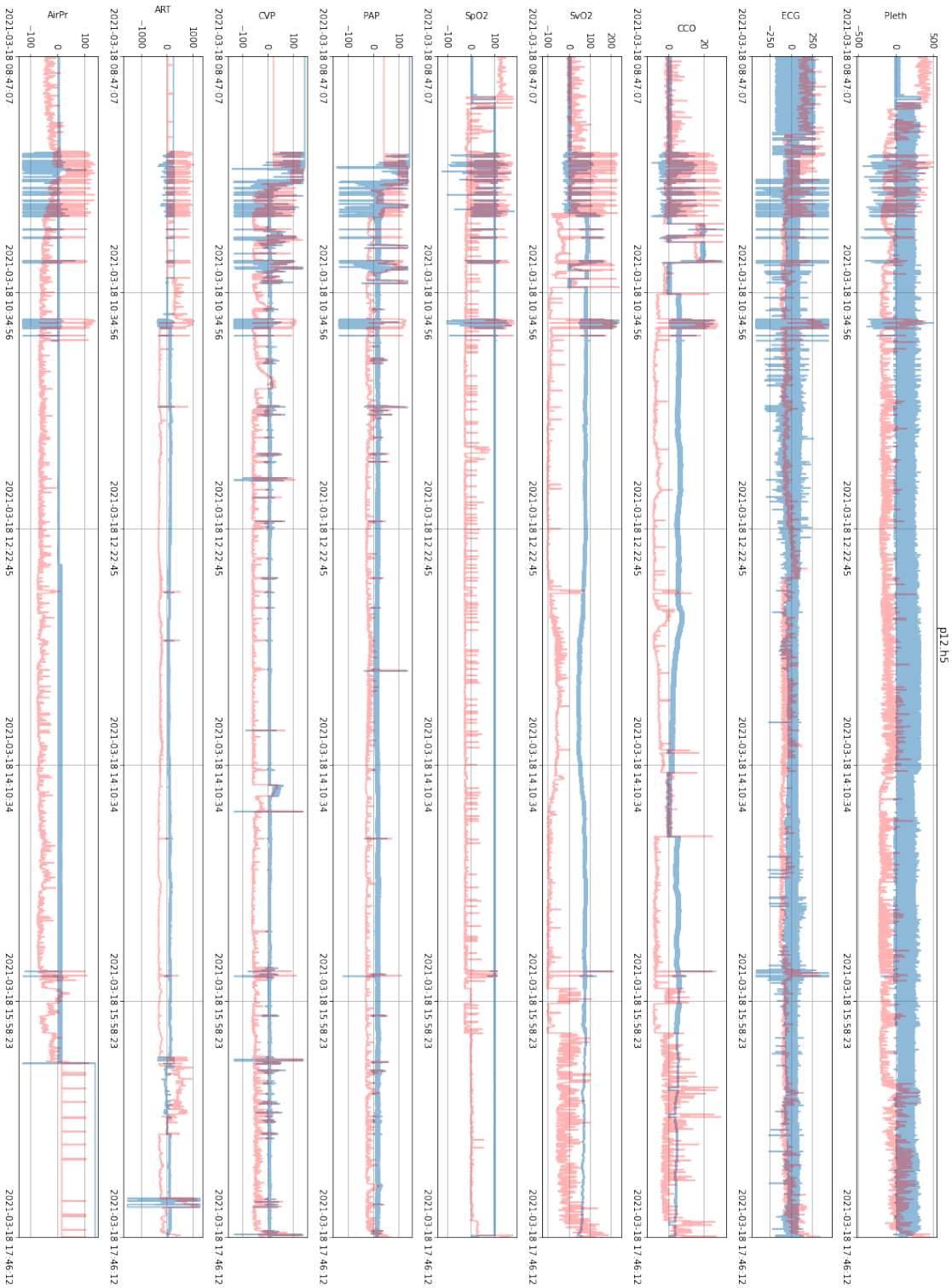

p07.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-09-24 08:42:10
2020-09-24 10:46:49
2020-09-24 12:51:28
2020-09-24 14:56:06
2020-09-24 17:00:45
2020-09-24 19:05:24

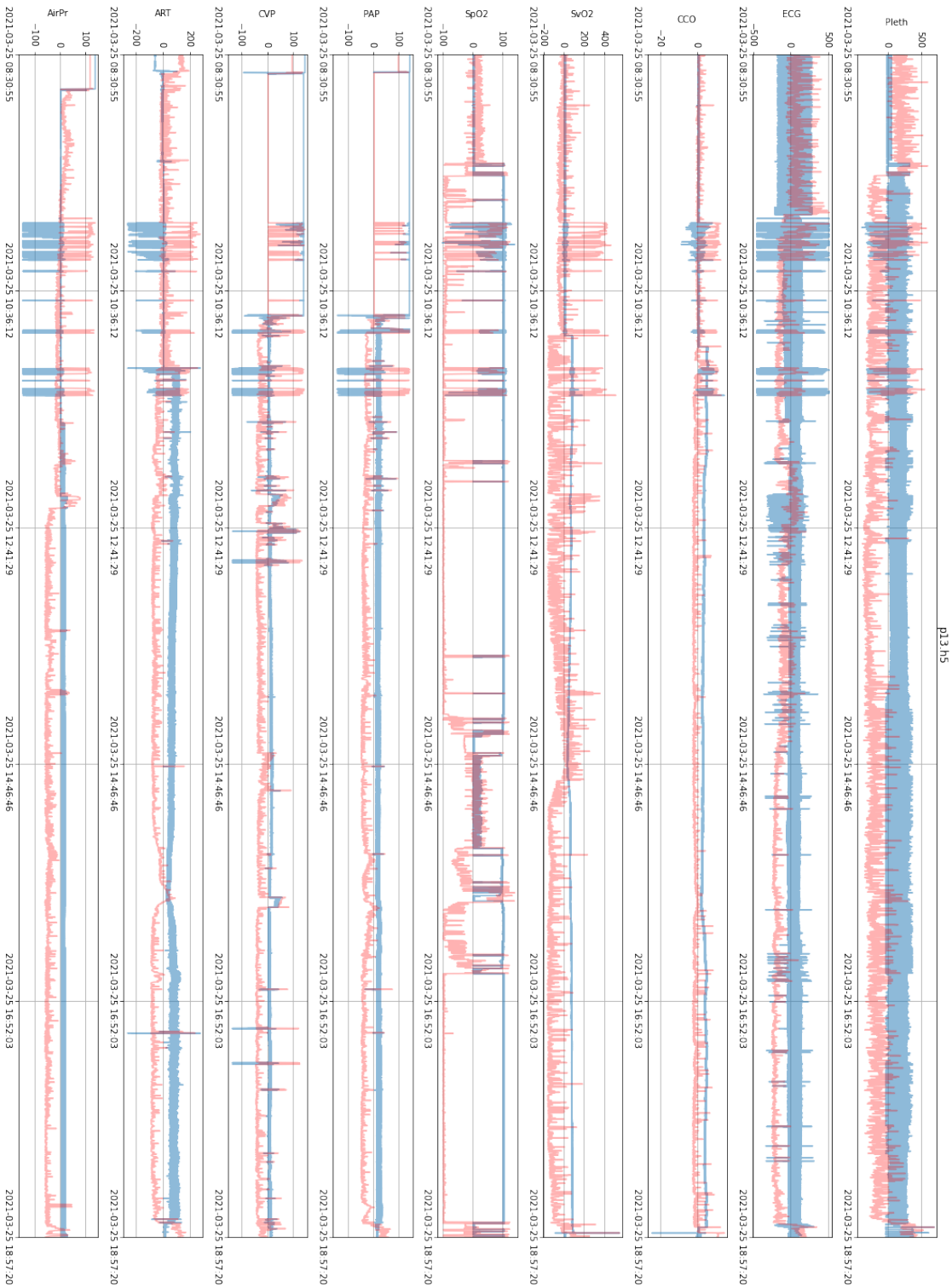

p08 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-10-13 08:01:57
2020-10-13 10:15:39
2020-10-13 12:29:21
2020-10-13 14:43:04
2020-10-13 16:56:46
2020-10-13 19:10:28

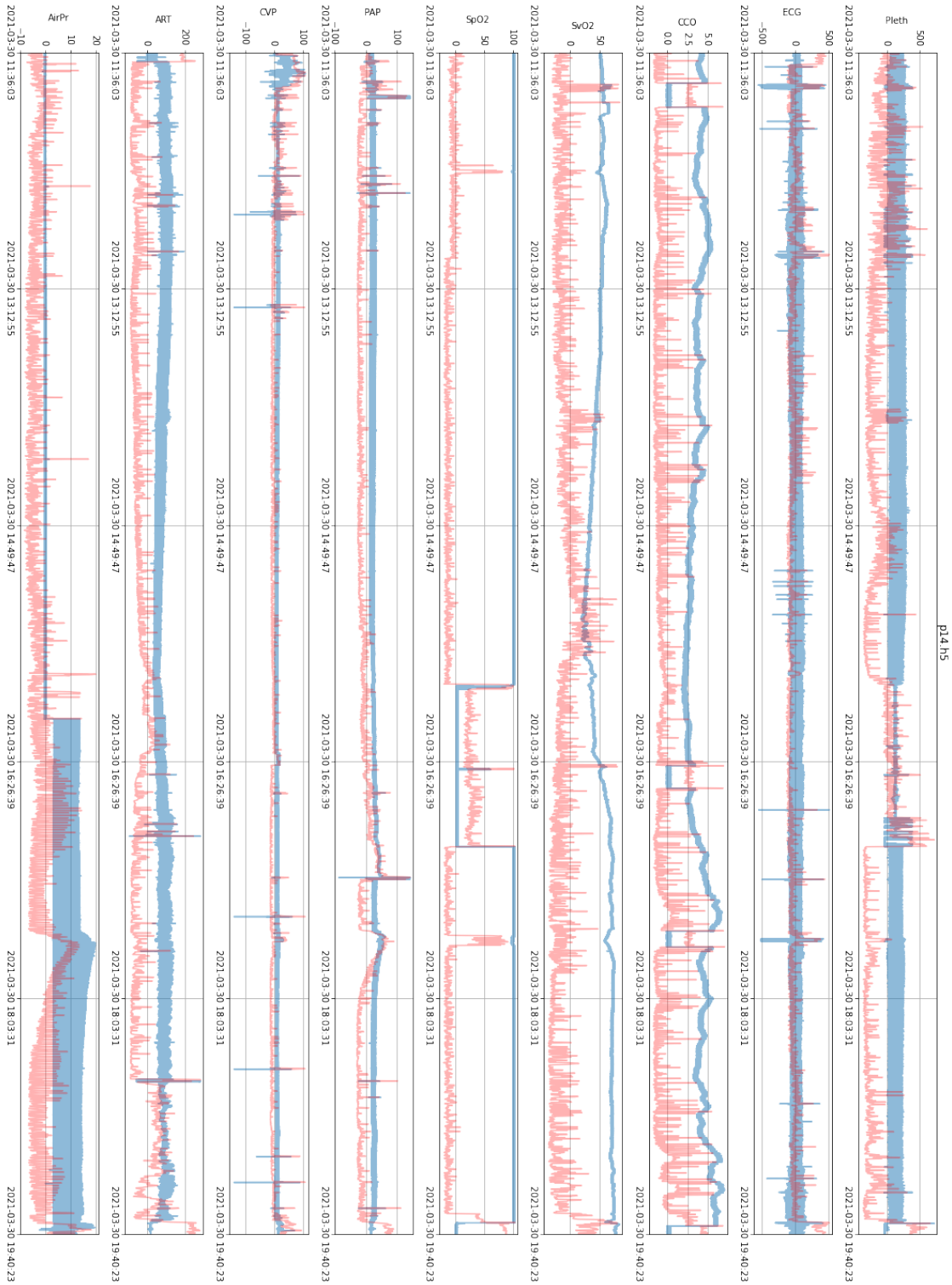

p09 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-10-20 09:21:38
2020-10-20 11:17:39
2020-10-20 13:13:41
2020-10-20 15:09:42
2020-10-20 17:05:44
2020-10-20 19:01:45

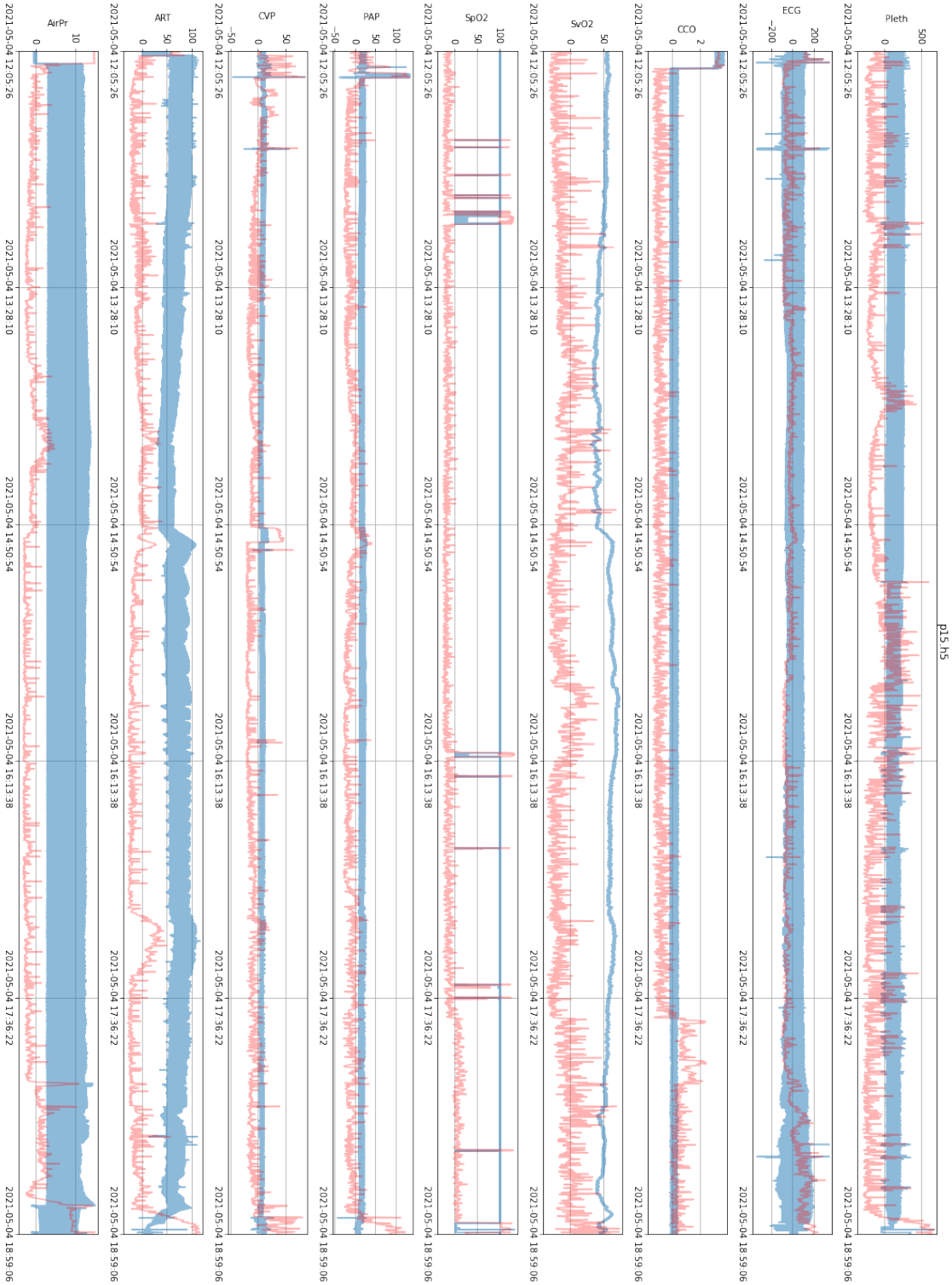

p10.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2020-10-27 11:23:27
2020-10-27 13:10:36
2020-10-27 14:57:46
2020-10-27 16:44:55
2020-10-27 18:32:04
2020-10-27 20:19:14

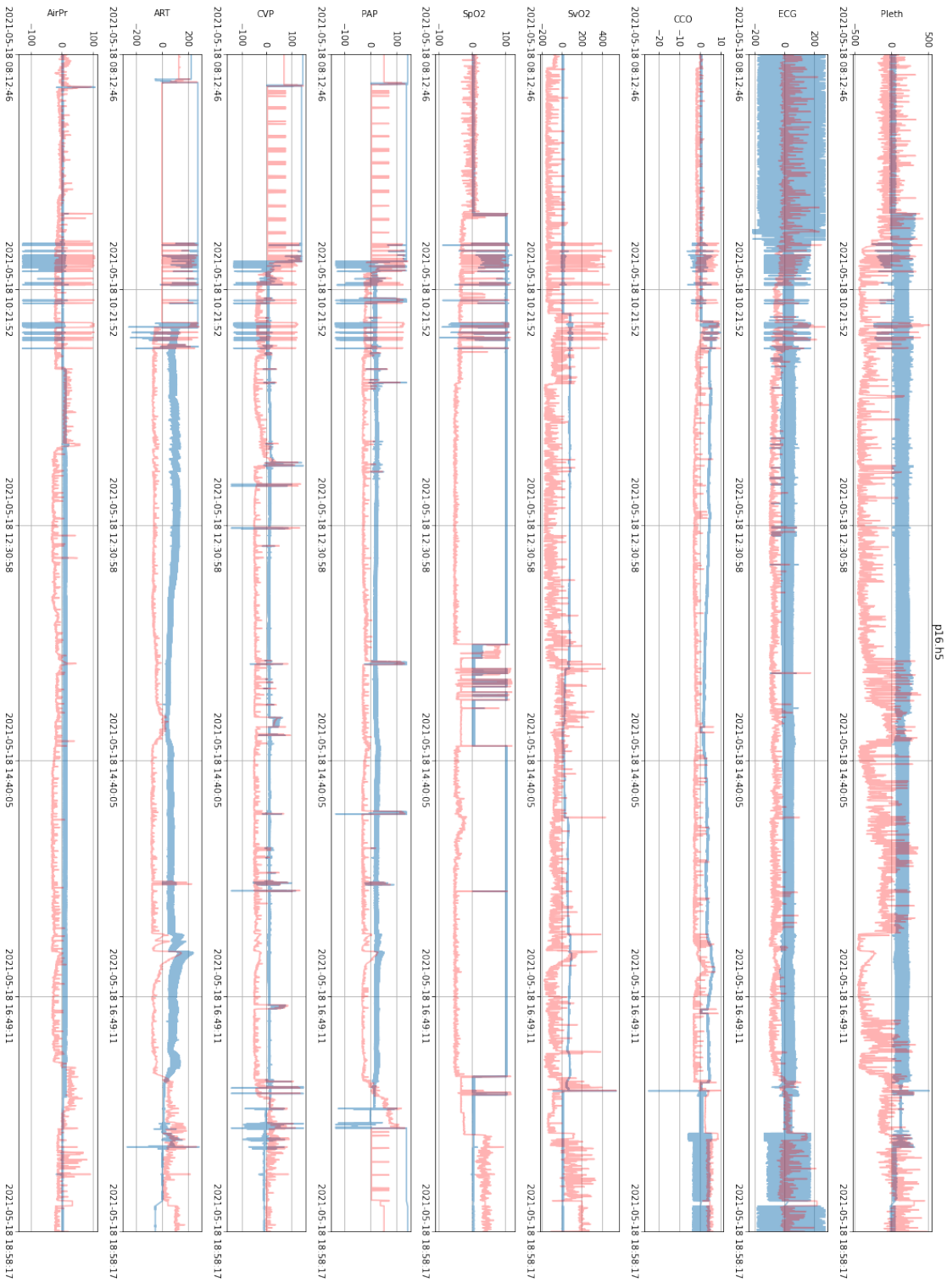

p11.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-03-16 12:48:15
2021-03-16 14:26:20
2021-03-16 16:04:26
2021-03-16 17:42:31
2021-03-16 19:20:36
2021-03-16 20:58:42

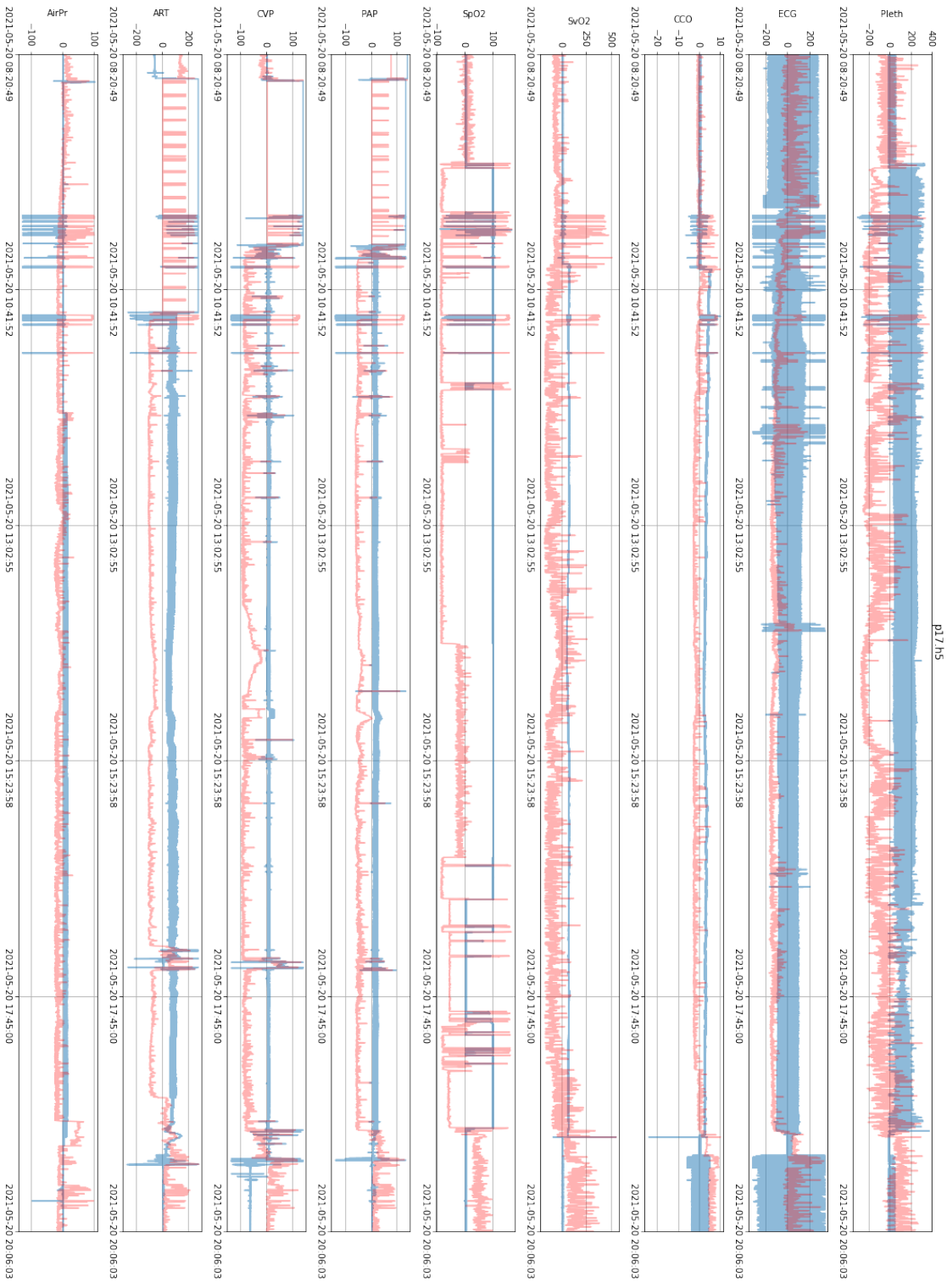

p12 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-03-18 08:47:07
2021-03-18 10:34:56
2021-03-18 12:22:45
2021-03-18 14:10:34
2021-03-18 15:58:23
2021-03-18 17:46:12

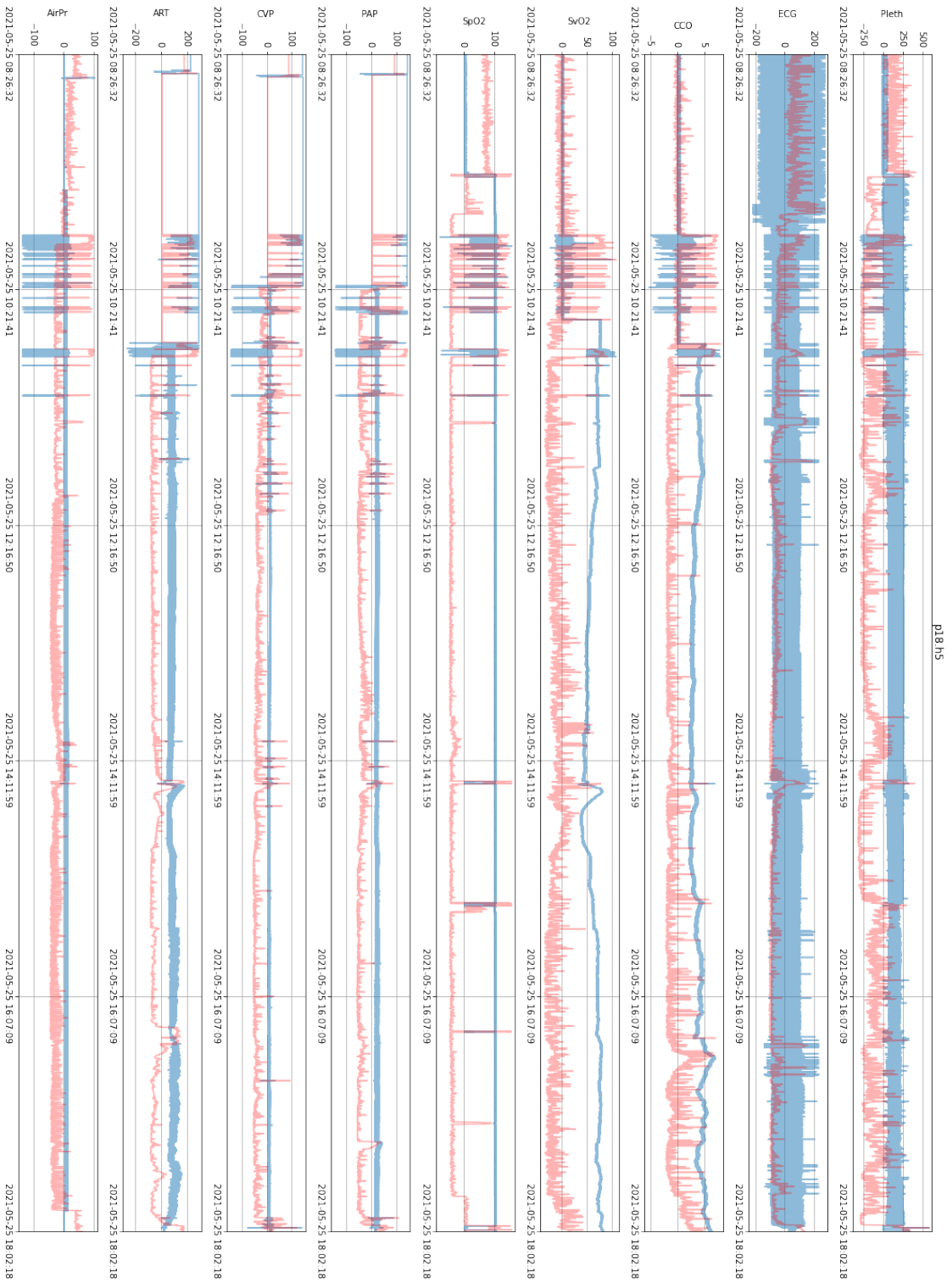

p13.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-03-25 08:30:55
2021-03-25 10:36:12
2021-03-25 12:41:29
2021-03-25 14:46:46
2021-03-25 16:52:03
2021-03-25 18:57:20

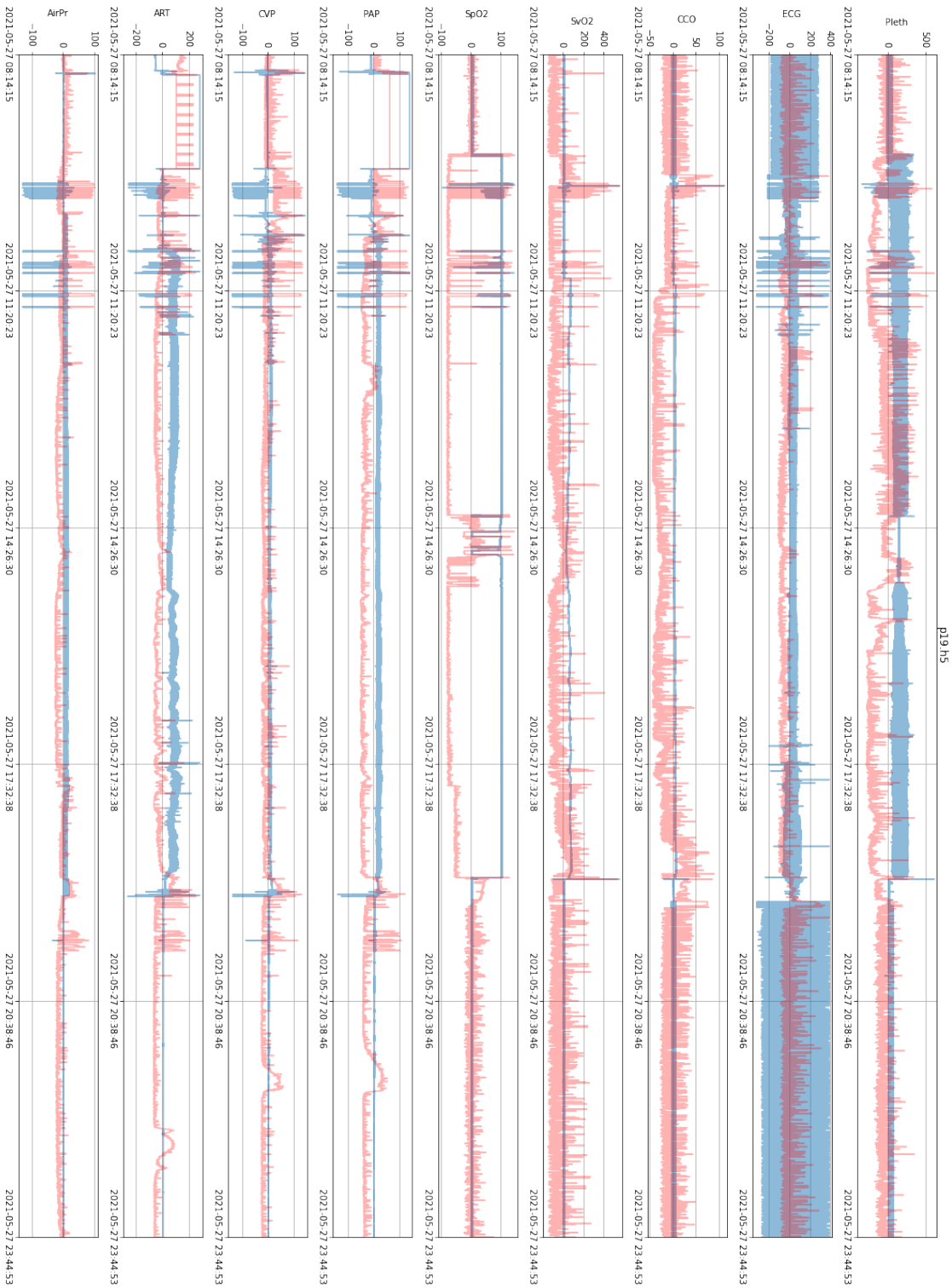

p14 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-03-30 11:36:03
2021-03-30 13:12:55
2021-03-30 14:49:47
2021-03-30 16:26:39
2021-03-30 18:03:31
2021-03-30 19:40:23

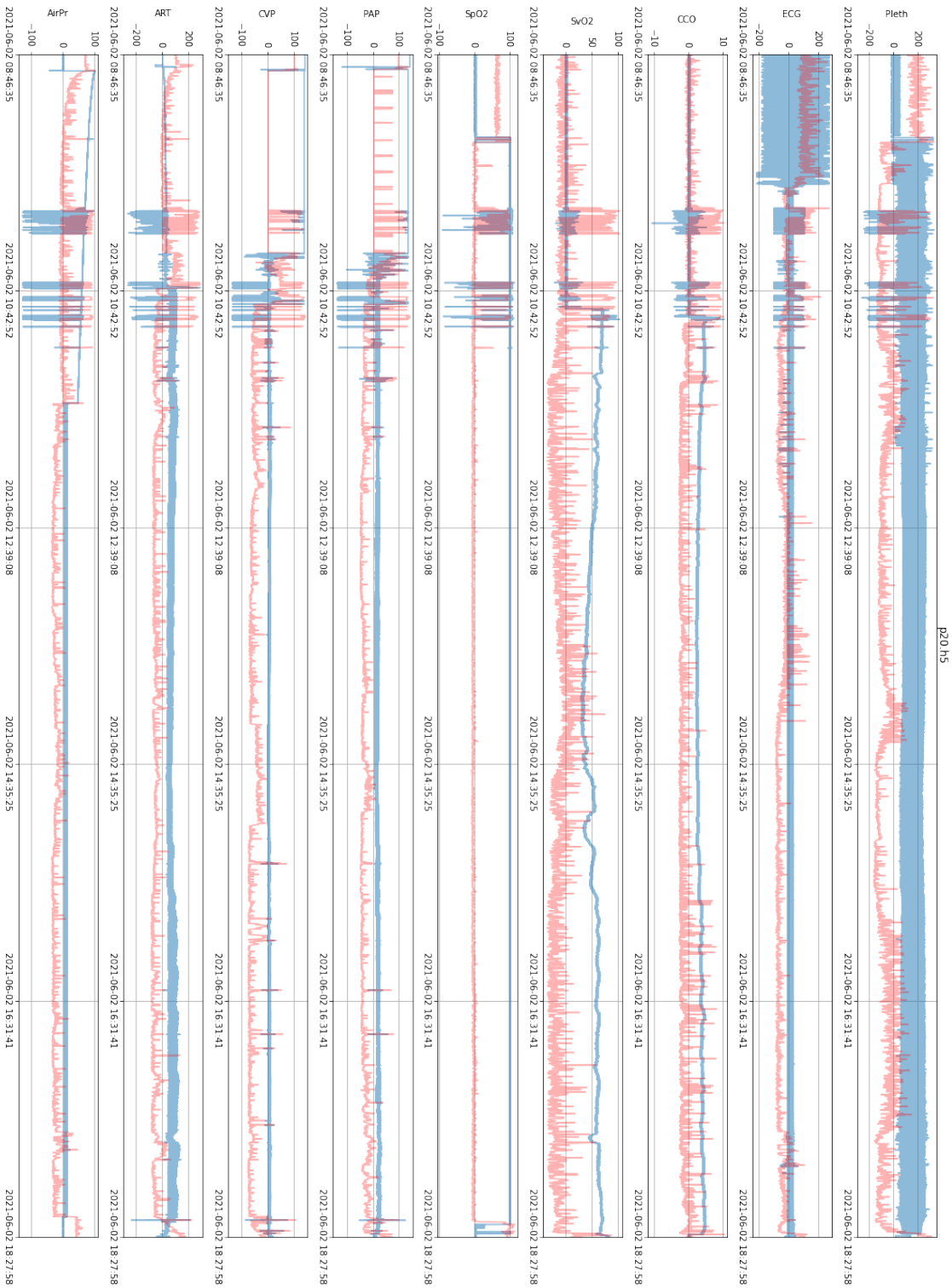

p15 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-05-04 12:05:26
2021-05-04 13:28:10
2021-05-04 14:50:54
2021-05-04 16:13:38
2021-05-04 17:36:22
2021-05-04 18:59:06

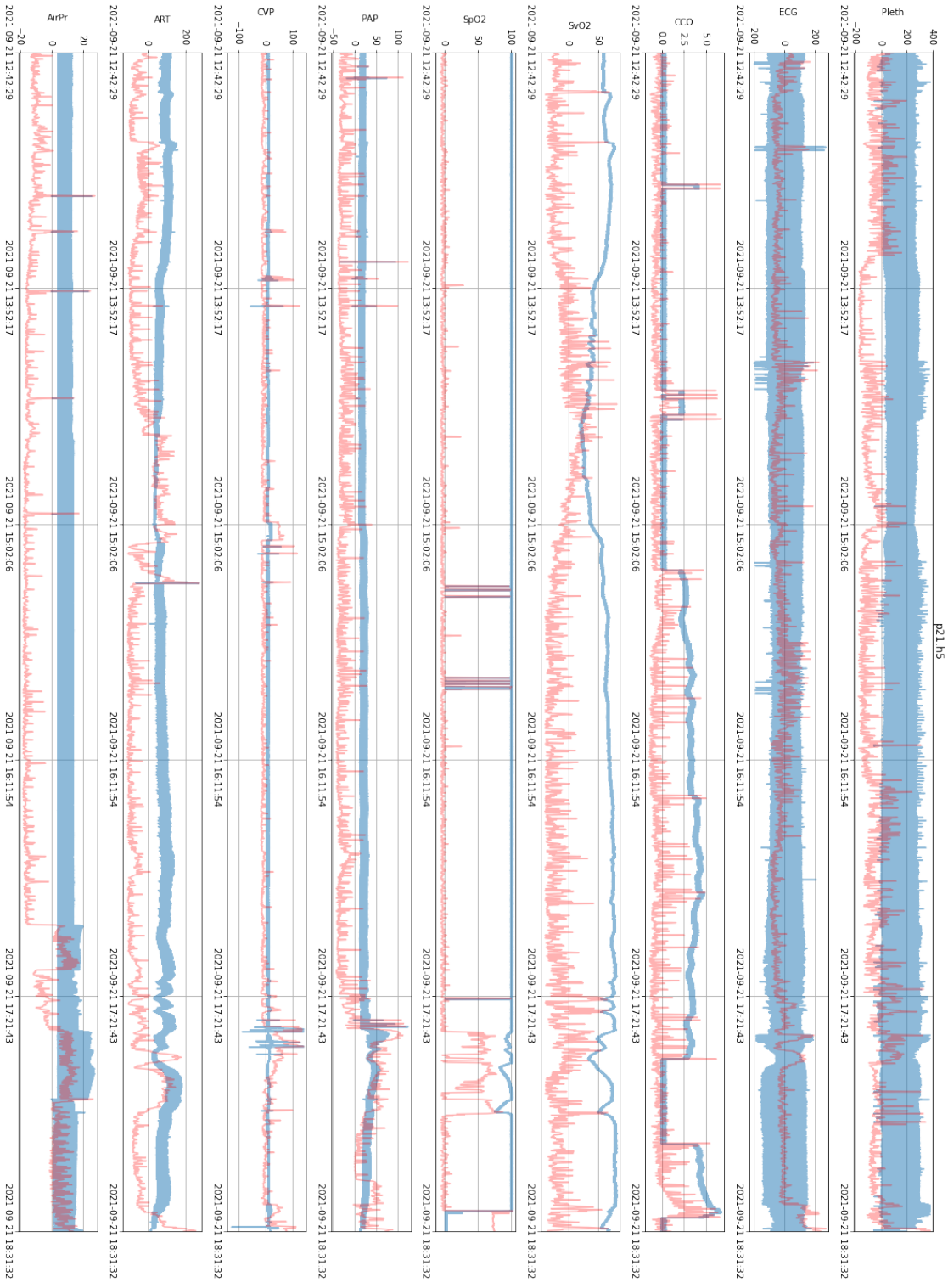

p16 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-05-18 08:12:46
2021-05-18 10:21:52
2021-05-18 12:30:58
2021-05-18 14:40:05
2021-05-18 16:49:11
2021-05-18 18:58:17

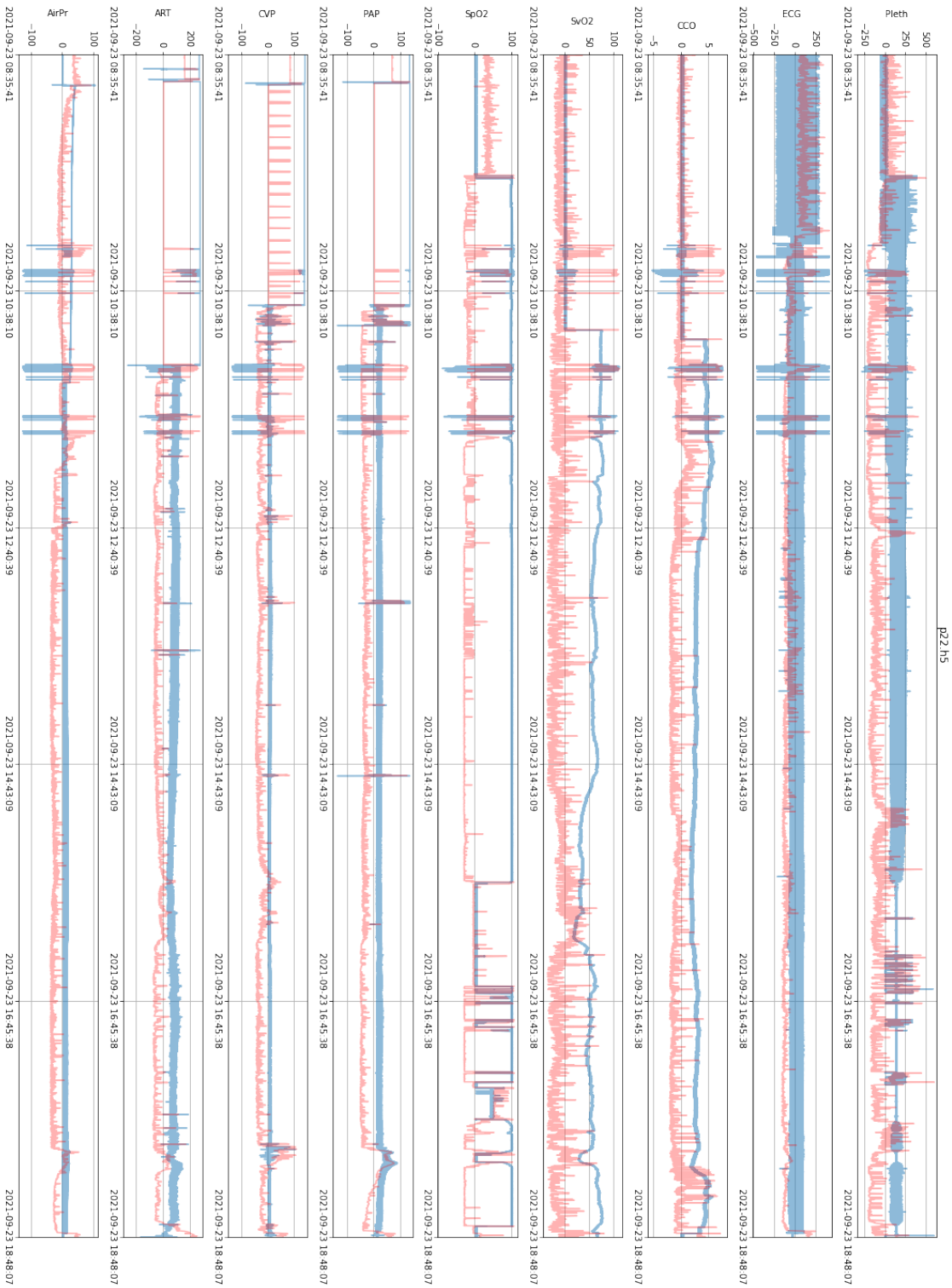

p17.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-05-20 08:20:49
2021-05-20 10:41:52
2021-05-20 13:02:55
2021-05-20 15:23:58
2021-05-20 17:45:00
2021-05-20 20:06:03

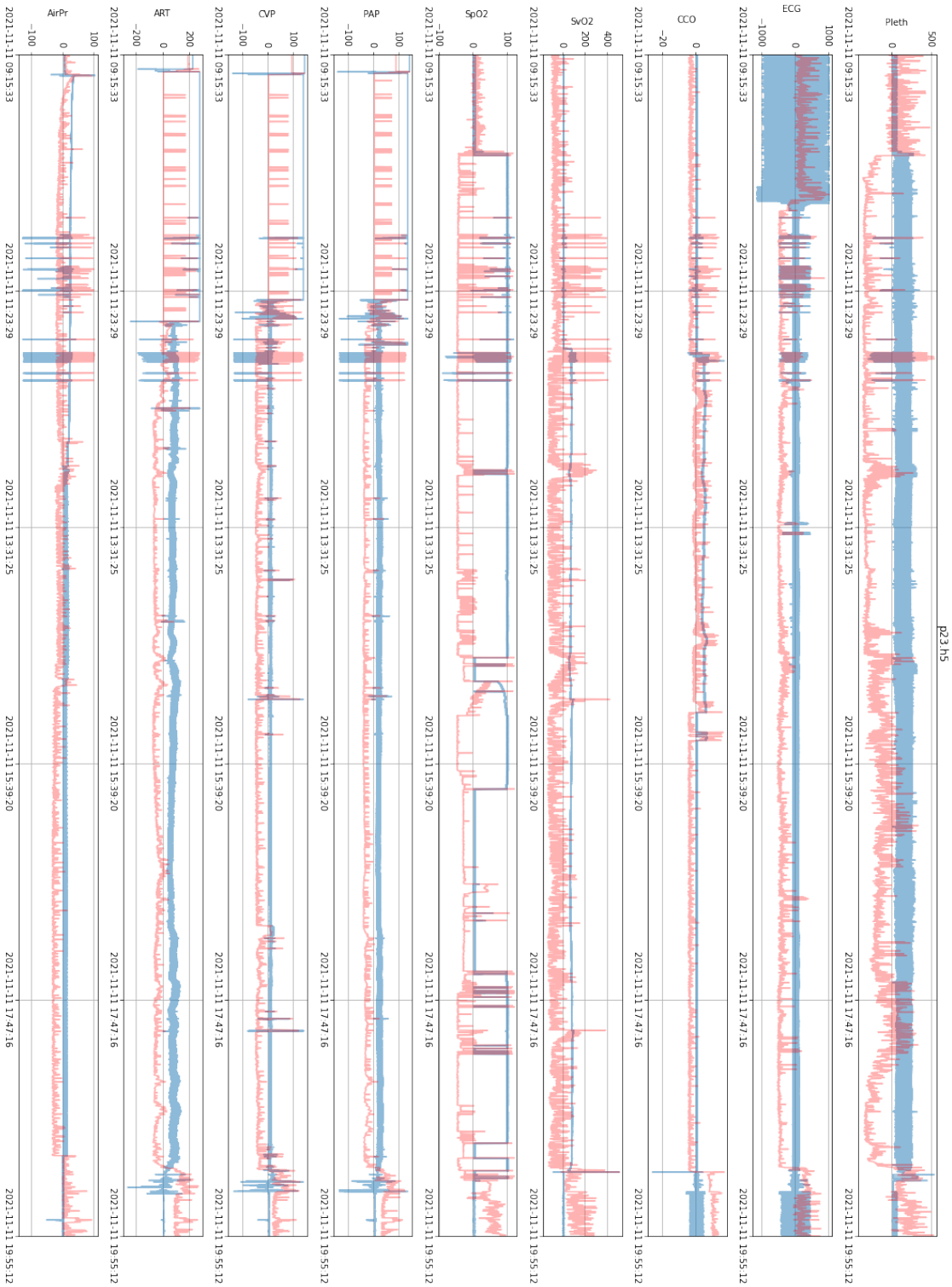

p18 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-05-25 08:26:32
2021-05-25 10:21:41
2021-05-25 12:16:50
2021-05-25 14:11:59
2021-05-25 16:07:09
2021-05-25 18:02:18

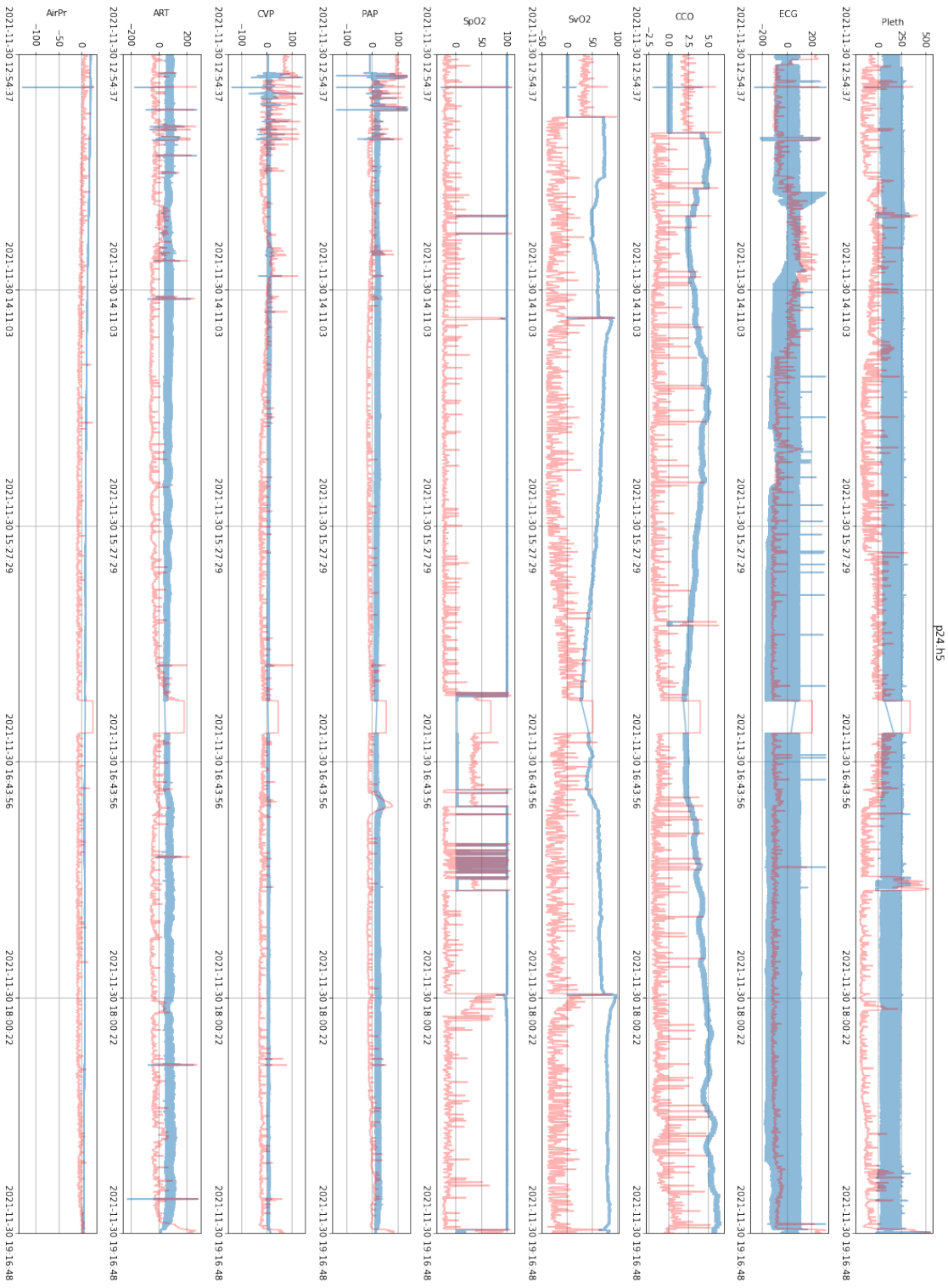

p19 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-05-27 08:14:15
2021-05-27 11:20:23
2021-05-27 14:26:30
2021-05-27 17:32:38
2021-05-27 20:38:46
2021-05-27 23:44:53

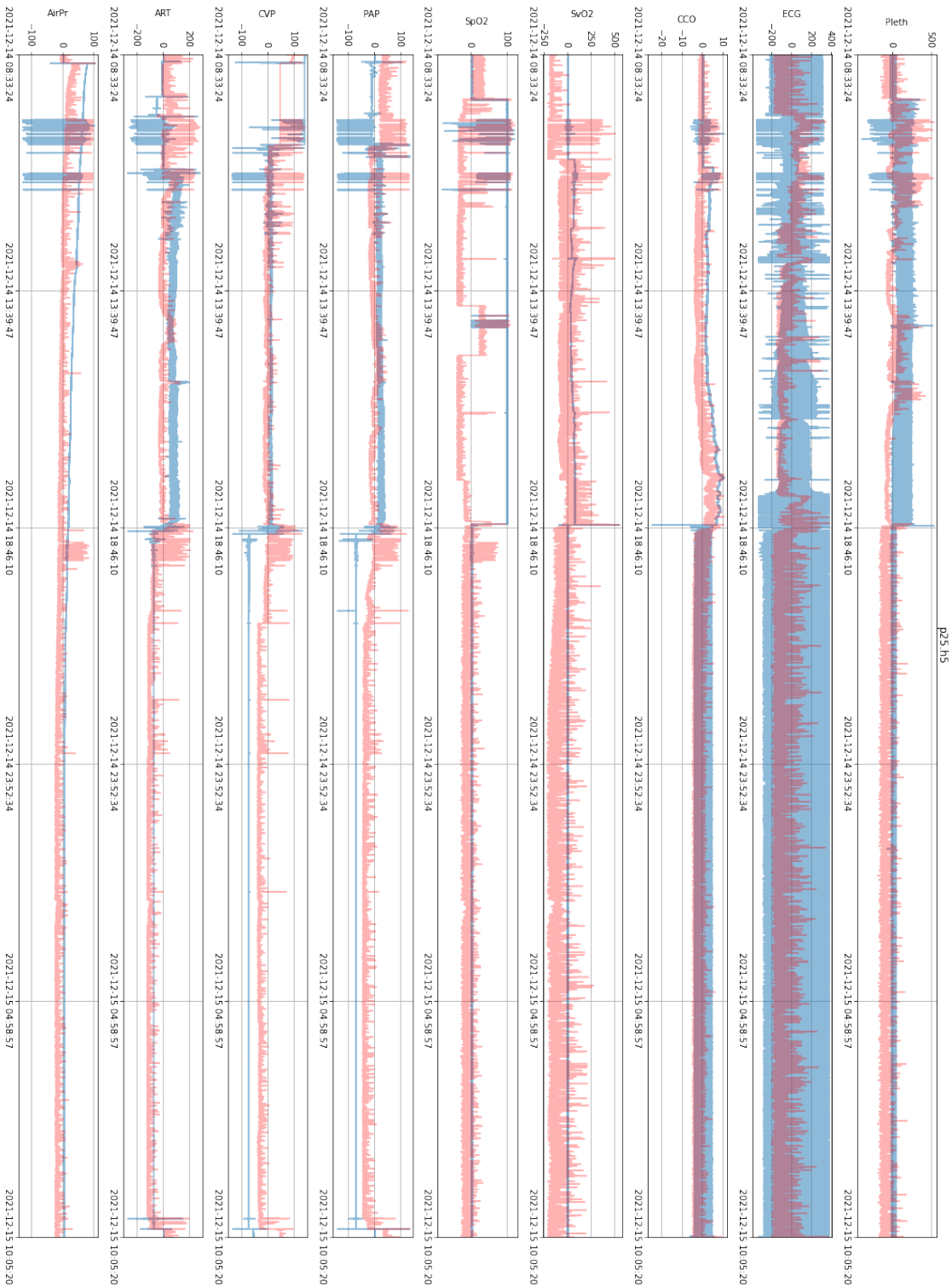

p20 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-06-02 08:46:35
2021-06-02 10:42:52
2021-06-02 12:39:08
2021-06-02 14:35:25
2021-06-02 16:31:41
2021-06-02 18:27:58


p21.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-09-21 12:42:29
2021-09-21 13:52:17
2021-09-21 15:02:06
2021-09-21 16:11:54
2021-09-21 17:21:43
2021-09-21 18:31:32


p22.h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-09-23 08:35:41
2021-09-23 10:38:10
2021-09-23 12:40:39
2021-09-23 14:43:09
2021-09-23 16:45:38
2021-09-23 18:48:07


p23 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-11-11 09:15:33
2021-11-11 11:23:29
2021-11-11 13:31:25
2021-11-11 15:39:20
2021-11-11 17:47:16
2021-11-11 19:55:12


p24 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-11-30 12:54:37
2021-11-30 14:11:03
2021-11-30 15:27:29
2021-11-30 16:43:56
2021-11-30 18:00:22
2021-11-30 19:16:48


p25 h5
Pleth
ECG
CCO
SvO2
SpO2
PAP
CVP
ART
AirPr
2021-12-14 08:33:24
2021-12-14 13:39:47
2021-12-14 18:46:10
2021-12-14 23:52:34
2021-12-15 04:58:57
2021-12-15 10:05:20

Bibliography

[1] Charu C Aggarwal. An introduction to outlier analysis. In *Outlier analysis*, pages 1–34. Springer, 2017. 2.5

[2] Ali N Akansu, Richard A Haddad, and Paul A Haddad. *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic press, 2001. A.1.1

[3] Mikhled Alfaouri and Khaled Daqrouq. Ecg signal denoising by wavelet transform thresholding. *American Journal of applied sciences*, 5(3):276–281, 2008. 2.2

[4] Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*, pages 15–27. Springer, 2002. 2.5

[5] Joachim Behar, Julien Oster, Qiao Li, and Gari D Clifford. Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE transactions on biomedical engineering*, 60(6):1660–1666, 2013. 1, 2.1, 3.1.1

[6] R Bousseljot, D Kreiseler, and A Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. 1995. 2.1

[7] Matt B Brearley, Michael F Heaney, and Ian N Norton. Physiological responses of medical team members to a simulated emergency in tropical field conditions. *Prehospital and disaster medicine*, 28(2):139–144, 2013. 2.4

[8] Andrew Bruce and Hong-Ye Gao. *Applied wavelet analysis with S-plus*. Springer Science & Business Media, 1996. 2.2

[9] Robert L Burr and Marie J Cowan. Autoregressive spectral models of heart rate variability: Practical issues. *Journal of Electrocardiology*, 25:224–233, 1992. A.1.1

[10] Pedro Miguel Caridade Gomes. *Development of an open-source Python toolbox for heart rate variability (HRV)*. PhD thesis, Hochschule für angewandte Wissenschaften Hamburg, 2019. 3.1.1

[11] Hsin-Tien Chiang, Yi-Yen Hsieh, Szu-Wei Fu, Kuo-Hsuan Hung, Yu Tsao, and Shao-Yi Chien. Noise reduction in ecg signals using fully convolutional denoising

autoencoders. *Ieee Access*, 7:60806–60813, 2019. 2.2

[12] Gari D Clifford, Francisco Azuaje, and Patrick Mcsharry. Ecg statistics, noise, artifacts, and missing data. *Advanced methods and tools for ECG data analysis*, 6(1):18, 2006. 1.1, A.1.1

[13] Gari D Clifford, Francisco Azuaje, Patrick McSharry, et al. *Advanced methods and tools for ECG data analysis*, volume 10. Artech house Boston, 2006. 2.1

[14] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017. A.1.1, A.1.1

[15] GD Clifford, J Behar, Q Li, and Iead Rezek. Signal quality indices and data fusion for determining clinical acceptability of electrocardiograms. *Physiological measurement*, 33(9):1419, 2012. 2.1, 2.1, 3.1.1, A.1.1

[16] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995. 2.2, 2.2

[17] Barbara J Drew, Patricia Harris, Jessica K Zegre-Hemsey, Tina Mammone, Daniel Schindler, Rebeca Salas-Boni, Yong Bai, Adelita Tinoco, Quan Ding, and Xiao Hu. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloS one*, 9(10):e110274, 2014. 1

[18] Mohamed Elgendi. Optimal signal quality index for photoplethysmogram signals. *Bioengineering*, 3(4):21, 2016. 2.4

[19] Tiago H Falk, Martin Maier, et al. Ms-qi: A modulation spectrum-based ecg quality index for telehealth applications. *IEEE Transactions on Biomedical Engineering*, 63(8):1613–1622, 2014. A.1.1

[20] Zoran Fejzo and Hanoch Lev-Ari. Adaptive laguerre-lattice filters. *IEEE Transactions on Signal Processing*, 45(12):3006–3016, 1997. A.1.1

[21] Patrick Flandrin, Gabriel Rilling, and Paulo Goncalves. Empirical mode decomposition as a filter bank. *IEEE signal processing letters*, 11(2):112–114, 2004. 2.2

[22] Chufan Gao, Fabian Falck, Mononito Goswami, Anthony Wertz, Michael R Pinsky, and Artur Dubrawski. Detecting patterns of physiological response to hemodynamic stress via unsupervised deep learning. *arXiv preprint arXiv:1911.05121*, 2019. 1

[23] Chufan Gao, A Dubrawski, MR Pinsky, G Clermont, and JH Yoon. Identification and explanation of severity of bleeding-induced hypovolemia using unsupervised deep learning. In *TP55. TP055 MECHANICAL VENTILATION,*

ICU MANAGEMENT, AND CV, pages A2841–A2841. American Thoracic Society, 2021. 1

[24] A Ghaffari, MR Homaeinezhad, M Akraminia, M Atarod, and M Daevaeiha. A robust wavelet-based multi-lead electrocardiogram delineation algorithm. *Medical engineering & physics*, 31(10):1219–1227, 2009. A.1.1

[25] Eduardo Gil, José María Vergara, and Pablo Laguna. Detection of decreases in the amplitude fluctuation of pulse photoplethysmography signal as indication of obstructive sleep apnea syndrome in children. *Biomedical Signal Processing and Control*, 3(3):267–277, 2008. 2.1, 2.4

[26] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000. 2.1, 2.3, 2.3, 2.3, A.1.1, A.2, A.2

[27] Mononito Goswami, Benedikt Boecking, and Artur Dubrawski. Weak supervision for affordable modeling of electrocardiogram data. In *AMIA Annual Symposium Proceedings*, volume 2021, page 536. American Medical Informatics Association, 2021. 1

[28] Patrick S Hamilton and Willis J Tompkins. Quantitative investigation of qrs detection rules using the mit/bih arrhythmia database. *IEEE transactions on biomedical engineering*, (12):1157–1165, 1986. 2.1

[29] Simon S Haykin. *Adaptive filter theory*. Pearson Education India, 2008. A.1.1

[30] Dieter Hayn, Bernhard Jammerbund, and Günter Schreier. Ecg quality assessment for patient empowerment in mhealth applications. In *2011 Computing in Cardiology*, pages 353–356. IEEE, 2011. A.1.1

[31] Dieter Hayn, Bernhard Jammerbund, and Günter Schreier. Qrs detection based ecg quality assessment. *Physiological measurement*, 33(9):1449, 2012. A.1.1

[32] Mohammad Reza Homaeinezhad, Ali Ghaffari, Mohammad Aghaee, Hamid Najjaran Toosi, and Reza Rahmani. A high-speed c++/mex solution for long-duration arterial blood pressure characteristic locations detection. *Biomedical Signal Processing and Control*, 7(2):151–172, 2012. A.1.1

[33] Borui Hou, Jianyong Yang, Pu Wang, and Ruqiang Yan. Lstm-based auto-encoder model for ecg arrhythmias classification. *IEEE Transactions on Instrumentation and Measurement*, 69(4):1232–1240, 2019. 2.2

[34] Marilyn Hravnak, Lujie Chen, Artur Dubrawski, Eliezer Bose, Gilles Clermont, and Michael R Pinsky. Real alerts and artifact classification in archived multi-signal vital sign monitoring data: implications for mining big data. *Journal of*

clinical monitoring and computing, 30(6):875–888, 2016. 1, 1

[35] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971):903–995, 1998. A.1.1

[36] Dorien Huysmans, Elena Smets, Walter De Raedt, Chris Van Hoof, Katleen Bogaerts, Ilse Van Diest, and Denis Helic. Unsupervised learning for mental stress detection-exploration of self-organizing maps. *Proc. of Biosignals 2018*, 4:26–35, 2018. A.1.1

[37] Franc Jager, Alessandro Taddei, George B Moody, Michele Emdin, G Antolič, Roman Dorn, Ales Smrdel, Carlo Marchesi, and Roger G Mark. Long-term st database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. *Medical and Biological Engineering and Computing*, 41(2):172–182, 2003. A.1.1

[38] Irena Jekova, Vessela Krasteva, Ivaylo Christov, and Roger Abächerli. Threshold-based system for noise detection in multilead ecg recordings. *Physiological measurement*, 33(9):1463, 2012. A.1.1

[39] Lars Johannesen. Assessment of ecg quality on an android platform. In *2011 Computing in Cardiology*, pages 433–436. IEEE, 2011. A.1.1

[40] Lars Johannesen and Loriano Galeotti. Automatic ecg quality scoring methodology: mimicking human annotators. *Physiological measurement*, 33(9):1479, 2012. A.1.1

[41] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and R Mark IV. Mimic-iv (version 0.4). *PhysioNet*, 2020. A.2

[42] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. A.2

[43] Johnteslade. Frequency domain representation of the dwt (discrete wavelet transform)., 2005. URL https://commons.wikimedia.org/wiki/File:Wavelets_-_DWT_Freq.png. (document), 2.2

[44] Johnteslade. Discrete wavelet transform - a 3 level filter bank, 2020. URL https://commons.wikimedia.org/wiki/File:Wavelets_-_Filter_Bank.png. (document), 2.1

[45] Sarang L Joshi, Rambabu A Vatti, and Rupali V Tornekar. A survey on ecg sig-

nal denoising techniques. In *2013 International Conference on Communication Systems and Network Technologies*, pages 60–64. IEEE, 2013. 1.1

[46] Antti Koski. Modelling ecg signals with hidden markov models. *Artificial intelligence in medicine*, 8(5):453–471, 1996. 1

[47] Jakub Kužílek, Michal Huptych, Václav Chudáček, Jiří Spilka, and Lenka Lhotská. Data driven approach to ecg signal quality assessment using multistep svm classification. In *2011 Computing in Cardiology*, pages 453–455. IEEE, 2011. A.1.1

[48] Philip Langley, Luigi Yuri Di Marco, Susan King, David Duncan, Costanzo Di Maria, Wenfeng Duan, Marjan Bojarnejad, Dingchang Zheng, John Allen, and Alan Murray. An algorithm for assessment of quality of ecgs acquired via mobile telephones. In *2011 Computing in Cardiology*, pages 281–284. IEEE, 2011. A.1.1

[49] Riccardo Leardi, R Boggia, and M Terrile. Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, 6(5):267–281, 1992. A.1.1

[50] Jinseok Lee, David D McManus, Sneh Merchant, and Ki H Chon. Automatic motion and noise artifact detection in holter ecg data using empirical mode decomposition and statistical approaches. *IEEE Transactions on Biomedical Engineering*, 59(6):1499–1506, 2011. A.1.1

[51] Chavdar Levkov, Georgy Mihov, Ratcho Ivanov, Ivan Daskalov, Ivaylo Christov, and Ivan Dotsinsky. Removal of power-line interference from the ecg: a review of the subtraction procedure. *BioMedical Engineering OnLine*, 4(1):1–18, 2005. A.1.1

[52] Q Li, RG Mark, and GD Clifford. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a kalman filter. *Physiological measurement*, 29(1):15, 2008. 2.1, 2.1, 3.1.1, A.1.1

[53] Qiao Li and G Clifford. Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments. *Computing in Cardiology*, 38:1, 2011. 2.1

[54] Qiao Li and Gari D Clifford. Signal quality and data fusion for false alarm reduction in the intensive care unit. *Journal of electrocardiology*, 45(6):596–603, 2012. 1, A.1.1

[55] Qiao Li, Cadathur Rajagopalan, and Gari D Clifford. A machine learning approach to multi-level ecg signal quality classification. *Computer methods and programs in biomedicine*, 117(3):435–447, 2014. 2.1, 3.1.1

[56] Xinyu Li, Michael R Pinsky, and Artur Dubrawski. Automated assessment of cardiovascular sufficiency using non-invasive physiological data. *Sensors*, 22(3):

1024, 2022. 1

[57] Chengyu Liu, Peng Li, Lina Zhao, Feifei Liu, and Ruxiang Wang. Real-time signal quality assessment for ecgs collected using mobile phones. In *2011 Computing in Cardiology*, pages 357–360. IEEE, 2011. A.1.1

[58] Chengyu Liu, Xiangyu Zhang, Lina Zhao, Feifei Liu, Xingwen Chen, Yingjia Yao, and Jianqing Li. Signal quality assessment and lightweight qrs detection for wearable ecg smartvest system. *IEEE Internet of Things Journal*, 6(2): 1363–1374, 2018. A.1.1

[59] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. 2.5

[60] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6 (1):1–39, 2012. 2.5

[61] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, pages 1–8, 2021. 1, 2.1, 2.1, 3.1.1, 3.2.3, A.2

[62] RG Mark, PS Schluter, G Moody, P Devlin, and D Chernoff. An annotated ecg database for evaluating arrhythmia detectors. In *IEEE Transactions on Biomedical Engineering*, volume 29, pages 600–600. IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC 345 E 47TH ST, NEW YORK, NY ..., 1982. 2.3

[63] Patrick E McSharry, Gari D Clifford, Lionel Tarassenko, and Leonard A Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE transactions on biomedical engineering*, 50(3):289–294, 2003. A.2

[64] Jonathan Moeyersons, Elena Smets, John Morales, Amalia Villa, Walter De Raedt, Dries Testelmans, Bertien Buyse, Chris Van Hoof, Rik Willems, Sabine Van Huffel, et al. Artefact detection and quality assessment of ambulatory ecg signals. *Computer methods and programs in biomedicine*, 182:105050, 2019. A.1.1

[65] Benjamin E Moody. Rule-based methods for ecg quality control. In *2011 Computing in Cardiology*, pages 361–363. IEEE, 2011. A.1.1

[66] George B Moody and Roger G Mark. The mit-bih arrhythmia database on cd-rom and software for use with it. In *[1990] Proceedings Computers in Cardiology*, pages 185–188. IEEE, 1990. 2.3

[67] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia

database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001. 1, 1, 2.1, 2.1, 2.2, 2.3, A.1.1, A.1.1

[68] George B Moody, W Muldrow, and Roger G Mark. A noise stress test for arrhythmia detectors. *Computers in cardiology*, 11(3):381–384, 1984. 1, 2.1, 2.1, 2.1, 2.2, 2.3, 3.2.1, A.1.1, A.1.1, A.1.1, A.1.1

[69] Chirag Nagpal, Xinyu Li, Michael R Pinsky, and Artur Dubrawski. Dynamically personalized detection of hemorrhage. In *Machine Learning for Healthcare Conference*, pages 109–123. PMLR, 2019. 1

[70] Rajaie Namas, Ali Ghuma, Andres Torres, Patricio Polanco, Hernando Gomez, Derek Barclay, Lisa Gordon, Sven Zenker, Hyung Kook Kim, Linda Hermus, et al. An adequately robust early tnf-α response is a hallmark of survival following trauma/hemorrhage. *PloS one*, 4(12):e8406, 2009. 1

[71] Hosein Naseri and MR Homaeinezhad. Detection and boundary identification of phonocardiogram sounds using an expert frequency-energy based metric. *Annals of biomedical engineering*, 41(2):279–292, 2013. A.1.1

[72] Hosein Naseri and MR Homaeinezhad. Electrocardiogram signal quality assessment using an artificially reconstructed target lead. *Computer Methods in Biomechanics and Biomedical Engineering*, 18(10):1126–1141, 2015. A.1.1

[73] Shamim Nemati, Atul Malhotra, and Gari D Clifford. Data fusion for improved respiration rate estimation. *EURASIP journal on advances in signal processing*, 2010:1–10, 2010. 2.1

[74] Christina Orphanidou, Timothy Bonnici, Peter Charlton, David Clifton, David Vallance, and Lionel Tarassenko. Signal-quality indices for the electrocardiogram and photoplethysmogram: Derivation and applications to wireless monitoring. *IEEE journal of biomedical and health informatics*, 19(3):832–838, 2014. 2.1, 3.1.1

[75] Jiapu Pan and Willis J Tompkins. A real-time qrs detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985. A.1.1, A.1.1

[76] Matteo Paoletti and Carlo Marchesi. Discovering dangerous patterns in long-term ambulatory ecg recordings using a fast qrs detection algorithm and explorative data analysis. *Computer Methods and programs in biomedicine*, 82(1): 20–30, 2006. A.1.1

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2.1

[78] Robi Polikar. The wavelet tutorial. part iv. multiresolution analysis: the discrete wavelet transform. *Rowan university*, 2008. 2.2

[79] Suranai Poungponsri and Xiao-Hua Yu. An adaptive filtering approach for electrocardiogram (ecg) signal noise reduction using neural networks. *Neurocomputing*, 117:206–213, 2013. 2.2

[80] Patrick X Quesnel, Adrian DC Chan, and Homer Yang. Real-time biosignal quality analysis of ambulatory ecg for detection of myocardial ischemia. In *2013 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5. IEEE, 2013. A.1.1

[81] Patrick X Quesnel, Adrian DC Chan, and Homer Yang. Signal quality and false myocardial ischemia alarms in ambulatory electrocardiograms. In *2014 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–5. IEEE, 2014. A.1.1

[82] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000. 2.5

[83] Stephen J Redmond, Nigel H Lovell, Jim Basilakis, and Branko G Celler. Ecg quality measures in telecare monitoring. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2869–2872. IEEE, 2008. A.1.1, A.1.1

[84] Stephen J Redmond, Yang Xie, D Chang, Jim Basilakis, and Nigel H Lovell. Electrocardiogram signal quality measures for unsupervised telehealth environments. *Physiological Measurement*, 33(9):1517, 2012. A.1.1

[85] Joshua S Richman and J Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 2000. 2.1, A.1.1

[86] Gabriel Rilling, Patrick Flandrin, Paulo Goncalves, et al. On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*, volume 3, pages 8–11. IEEER Grado, 2003. 2.2

[87] Mohammed Saeed, Christine Lieu, Greg Raber, and Roger G Mark. Mimic ii: a massive temporal icu patient database to support research in intelligent patient monitoring. In *Computers in cardiology*, pages 641–644. IEEE, 2002. 2.1

[88] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011. 2.1, A.1.1, A.1.1, A.2

[89] Reza Sameni, Christian Jutten, and Mohammad B Shamsollahi. Multichannel electrocardiogram decomposition using periodic component analysis. *IEEE transactions on biomedical engineering*, 55(8):1935–1940, 2008. 2.1

[90] Arun Kumar Sangaiah, Maheswari Arumugam, and Gui-Bin Bian. An intelligent learning approach for improving ecg signal classification and arrhythmia analysis. *Artificial intelligence in medicine*, 103:101788, 2020. A.1.1

[91] Udit Satija, Barathram Ramkumar, and M Sabarimalai Manikandan. A review of signal processing techniques for electrocardiogram signal quality assessment. *IEEE reviews in biomedical engineering*, 11:36–52, 2018. 1.1

[92] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001. 2.5

[93] Ikaro Silva. Physionet 2010 challenge: A robust multi-channel adaptive filtering approach to the estimation of physiological recordings. In *2010 Computing in Cardiology*, pages 313–316. IEEE, 2010. A.1.1

[94] Ikaro Silva, George B Moody, and Leo Celi. Improving the quality of ecgs collected using mobile phones: The physionet/computing in cardiology challenge 2011. In *2011 Computing in Cardiology*, pages 273–276. IEEE, 2011. 1, 2.1, 2.1, 2.1, 2.3, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1, A.1.1

[95] Ikaro Silva, Joon Lee, and Roger G Mark. Signal quality estimation with multichannel adaptive filtering in intensive care settings. *IEEE Transactions on Biomedical Engineering*, 59(9):2476–2485, 2012. A.1.1

[96] Brij N Singh and Arvind K Tiwari. Optimal selection of wavelet basis function applied to ecg signal denoising. *Digital signal processing*, 16(3):275–287, 2006. 2.2

[97] Angela Stallone, Antonio Cicone, and Massimo Materassi. New insights and best practices for the successful use of empirical mode decomposition, iterative filtering and derived algorithms. *Scientific reports*, 10(1):1–15, 2020. 2.2

[98] Thomas Ho Chee Tat, Chen Xiang, and Lim Eng Thiam. Physionet challenge 2011: improving the quality of electrocardiography data collected using real time qrs-complex and t-wave detection. In *2011 Computing in Cardiology*, pages 441–444. IEEE, 2011. A.1.1

[99] PE Tikkanen. Nonlinear wavelet and wavelet packet denoising of electrocardiogram signal. *Biological cybernetics*, 80(4):259–267, 1999. 2.2

[100] Michael Tipping. The relevance vector machine. *Advances in neural information processing systems*, 12, 1999. A.1.1

[101] Carolina Varon, Dries Testelmans, Bertien Buyse, Johan AK Suykens, and Sabine Van Huffel. Robust artefact detection in long-term ecg recordings based on autocorrelation function similarity and percentile analysis. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3151–3154. IEEE, 2012. A.1.1

[102] JY Wang. A new method for evaluating ecg signal quality for multi-lead arrhythmia analysis. In *Computers in Cardiology*, pages 85–88. IEEE, 2002. A.1.1

[103] Binwei Weng, Manuel Blanco-Velasco, and Kenneth E Barner. Ecg denoising based on the empirical mode decomposition. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1–4. IEEE, 2006. 2.2

[104] Henian Xia, Gabriel A Garcia, Joseph C McBride, Adam Sullivan, Thibaut De Bock, Jujhar Bains, Dale C Wortham, and Xiaopeng Zhao. Computer algorithms for evaluating the quality of ecgs in real time. In *2011 Computing in Cardiology*, pages 369–372. IEEE, 2011. A.1.1

[105] Zhidong Zhao and Yefei Zhang. Sqi quality evaluation mechanism of single-lead ecg signal based on simple heuristic fusion and fuzzy comprehensive evaluation. *Frontiers in Physiology*, 9:727, 2018. 2.1, 2.1, 3.1.1

[106] W Zong, GB Moody, and D Jiang. A robust open-source algorithm to detect onset and duration of qrs complexes. In *Computers in Cardiology, 2003*, pages 737–740. IEEE, 2003. 2.1